UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

ELIGIENDO ENTRE REGRESIÓN LOGÍSTICA Y ANÁLISIS DISCRIMINANTE

T E S I S

QUE PARA OBTENER EL TÍTULO DE

A C T U A R I O

P R E S E N T A :

JOSÉ CAMARILLO RODRÍGUEZ

DIRECTORA DE TESIS:

MAT. MARGARITA ELVIRA CHÁVEZ CANO

FACULTAD DE CIENCIAS
UNAM

2010

Hoja de Datos del Jurado

1.Datos del alumno Apellido paterno Apellido materno Nombre(s) Teléfono Universidad Nacional Autónoma de México Facultad de Ciencias Carrera Número de cuenta	1. Datos del alumno Camarillo Rodríguez José 57281000 Ext. 6021 Universidad Nacional Autónoma de México Facultad de Ciencias Actuaría 098513367
2. Datos del tutor Grado Nombre(s) Apellido paterno Apellido materno	2. Datos del tutor Mat. Margarita Elvira Chávez Cano
3. Datos del sinodal 1 Grado Nombre(s) Apellido paterno Apellido materno	3. Datos del sinodal 1 Dra. Ruth Selene Fuentes García
4. Datos del sinodal 2 Grado Nombre(s) Apellido paterno Apellido materno	4. Datos del sinodal 2 Act. Jaime Vázquez Alamilla
5. Datos del sinodal 3 Grado Nombre(s) Apellido paterno Apellido materno	5. Datos del sinodal 3 Act. Rosa Daniela Chávez Aguilar
6. Datos del sinodal 4 Grado Nombre(s) Apellido paterno Apellido materno	6. Datos del sinodal 4 M. en A.P. María del Pilar Alonso Reyes
7. Datos de la tesis. Título Número de páginas Año	7. Datos de la tesis Eligiendo entre regresión logística y análisis discriminante 71 2010

Índice general

Agradecimientos

A Dios, por permitirme llegar a este importante momento de mi vida.

A mi esposa e hija, por darme la inspiración que necesitaba para la conclusión de este trabajo. Gracias por aguantarme. Las amo.

A mis padres, a quienes debo lo que soy. Gracias por la confianza y apoyo incondicional durante mis años de estudio y por el sacrifico que hicieron para ayudarme a alcanzar este sueño.

A mis hermanos, Petty, Lily y Gerardo. En la distancia siempre estuvieron conmigo, su recuerdo nunca me hizo sentir solo y su cariño me motivó a seguir adelante.

A mis amigos de siempre: Lizeth, Nadia, Wendy, Ángeles, Miguel, Ambrosio y Alberto, por los inolvidables momentos que compartieron conmigo y por su incondicional apoyo en todos estos años. También quiero agradecer a mi amigo y maestro Luis Alberto Vázquez Maison, por facilitarme las herramientas que necesitaba cuando empecé a escribir esta tesis y por sus consejos y desinteresada ayuda.

A mi directora de tesis Mat. Margarita Chávez Cano, por su asesoría y acertados consejos. Gracias por su tiempo y paciencia.

A mis sinodales, Dra. Ruth Fuentes García, Mtra. Ma. del Pilar Alonso Reyes, Act. Daniela Chávez Aguilar y Act. Jaime Vázquez Alamilla, por el tiempo dedicado a este trabajo y por sus valiosos comentarios y sugerencias.

A la Universidad Nacional Autónoma de México y en especial a la Facultad de Ciencias, por darme la oportunidad de formarme entre sus aulas.

Introducción

El problema de la clasificación es uno de los temas que aparece con más frecuencia en la actividad científica y constituye un proceso en casi cualquier actividad humana, de tal manera que en la resolución de problemas y en la toma de decisiones, la primera parte de la tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología co-rrespondiente. Por ejemplo en medicina, ciencia en la que el diagnóstico es una parte fundamental, siendo una fase previa para la aplicación de un tratamiento, diagnosticar equivale a clasificar un sujeto en una patología concreta con base en los datos correspondientes a un estudio previo y observaciones complementarias.

Desde el punto de vista estadístico podemos ver a la clasificación como un problema en el que se trata de determinar un criterio para etiquetar un individuo como perteneciente a alguno de varios grupos, en los que éstos están bien definidos. Este criterio se construye a partir de los valores de una serie limitada de parámetros. En este caso, la técnica más utilizada se conoce con el nombre de análisis discriminante, aunque existen otras opciones como es el uso de la regresión logística.

El análisis discriminante estudia las técnicas de clasificación de sujetos en grupos ya definidos, a partir de una muestra de n sujetos en los que se ha medido p variables cuantitativas independientes, las cuales se utilizarán para tomar la decisión en cuanto al grupo en el que se clasifica cada sujeto, a través del modelo matemático estimado a partir de los datos.

Mediante las ecuaciones estimadas en el procedimiento de análisis discriminante se obtiene un mecanismo para asignar un sujeto a uno de los grupos, a partir de los valores de las variables explicativas.

El principal inconveniente del análisis discriminante tradicional radica en que supone que los grupos pertenecen a poblaciones con distribución de probabilidad normal multivariada para las variables explicativas $(X_1, X_2, ...X_p)$, con igual matriz de varianzas y covarianzas. Por ello no debiera incluirse en el modelo variables que no cumplieran esa condición, lo que no permite por ejemplo la utilización de variables cualitativas.

Sin embargo, en el modelo de regresión logística, se estima la probabilidad de un suceso en función de un conjunto de variables explicativas y en la construcción del mismo no hay ninguna suposición en cuanto a la distribución de probabilidad de esas variables, por lo que pueden intervenir variables no normales y variables cualitativas. Si tenemos dos grupos, de tal manera que un sujeto pertenece al grupo I o al II (por ejemplo tiene hipertensión o no la tiene), podemos considerar el modelo de regresión logística como una fórmula para calcular la probabilidad de pertenecer a uno de esos grupos, y estimar así la probabilidad de que una observación X pertenezca al grupo I, o su complementaria la probabilidad de que pertenezca al grupo II. De esta forma, podemos considerar la regresión logística como una alternativa al análisis discriminante.

Si los dos grupos son normales con igual matriz de covarianzas, el análisis discriminante es preferido sobre la regresión logística. Si el supuesto de normalidad es violado (como a menudo sucede), preferimos la regresión logística para resolver el problema.

Esta tesis está estructurada en tres capítulos. En el capítulo uno estableceremos las consideraciones metodológicas de la regresión logística. Desarrollaremos y describiremos de manera detallada la construcción de los modelos simple y múltiple.

En el capítulo dos revisaremos la teoría del análisis discriminante. Primero

daremos una explicación general de la idea que hay detrás de esta técnica y posteriormente haremos el desarrollo teórico.

Finalmente, en el capítulo tres realizaremos una aplicación de las dos técnicas a un conjunto de datos y compararemos los resultados con base en el número de clasificaciones correctas que tiene cada una.

Capítulo 1

Regresión logística

1.1. Introducción

Los métodos de regresión se han convertido en un componente integral en cualquier análisis asociado con describir la relación entre una variable de respuesta y una o más variables explicativas. A menudo la variable de respuesta es discreta y toma dos o más posibles valores. En los últimos años el modelo de regresión logística se ha convertido, en muchas áreas, el método estándar de análisis en esta situación.

Antes de comenzar con el estudio de la regresión logística, es importante entender que el objetivo de un análisis usando este método es el mismo que el de cualquier otra técnica de modelación usada en estadística: encontrar el modelo que mejor describa la relación entre una variable de respuesta (también conocida como variable dependiente) y un conjunto de variables independientes, llamadas a menudo covariables. El ejemplo más común para modelar esta situación es el modelo de regresión lineal, donde se supone que la variable de respuesta es continua.

Lo que distingue a la regresión logística del modelo de regresión lineal es que la variable de respuesta en la regresión logística es binaria o dicotómica. Esta diferencia se refleja en la elección del modelo paramétrico y en las hipótesis que se hacen en ambos casos. Una vez que esta diferencia ha sido

considerada, los métodos utilizados en el análisis de regresión logística siguen los mismos principios usados en regresión lineal.

En cualquier problema de regresión, una medida muy importante es el valor esperado de la variable de respuesta dado el valor de la variable independiente. Esta cantidad se conoce como la esperanza condicional y se expresa como $E(Y \mid x)$, donde Y representa la variable de respuesta y x el valor de la variable independiente. La cantidad $E(Y \mid x)$ se lee como "el valor esperado de Y, dado el valor de x". En regresión lineal suponemos que esta esperanza puede ser expresada como una ecuación lineal en x, como:

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

Esta expresión implica que $E(Y \mid x)$ al igual que x, puede tomar cualquier valor entre $-\infty$ y ∞.

Con datos binarios, esta esperanza condicional debe ser mayor o igual que cero y menor o igual que uno, es decir, $0 \leq E(Y \mid x) \leq 1$.

Se han propuesto varias distribuciones para el análisis de datos binarios, pero hay dos razones por las cuales se elige la distribución logística: (1) desde el punto de vista matemático es una función muy sencilla y flexible, y (2) se presta para una fácil interpretación.

Para simplificar notación, cuando utilicemos la distribución logística, usaremos $\pi(x)$=$E(Y \mid x)$ para representar la esperanza condicional de Y dado x. La forma específica del modelo de regresión logística que utilizaremos es:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{1.1}$$

Una transformación que será de suma importancia en el estudio de regresión logística es la transformación *logit*. Esta transformación está definida en términos de $\pi(x)$ como sigue:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

La importancia de esta transformación es que $g(x)$ tiene muchas de las propiedades del modelo de regresión lineal: $g(x)$ es lineal en sus parámetros, puede ser continua y tomar valores entre $-\infty$ y ∞, dependiendo del rango de x.

Una diferencia importante entre los modelos de regresión logística y lineal es la distribución condicional de la variable de respuesta. En el modelo de regresión lineal suponemos que una observación de la variable de respuesta puede ser expresada como $y = E(Y \mid x) + \varepsilon$. La cantidad ε se conoce como el error y expresa la distancia entre la observación y la media condicional. La hipótesis más común que se hace sobre ε es que tiene una distribución normal con media cero y varianza constante. De aquí, se sigue que la distribución condicional de la variable de respuesta y dado x es normal con media $E(Y \mid x)$ y varianza constante. Sin embargo, éste no es el caso cuando la variable de respuesta es dicotómica. En esta situación podemos expresar el valor de la variable de respuesta dado x como $y = \pi(x) + \varepsilon$. Aquí la cantidad ε puede tomar uno de dos posibles valores. Si $y = 1$ entonces $\varepsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$ y si $y = 0$ entonces $\varepsilon = -\pi(x)$ con probabilidad $1 - \pi(x)$. Por lo tanto ε tiene una distribución normal con media cero y varianza $\pi(x)[1 - \pi(x)]$. Esto es, la distribución condicional de la variable de respuesta es binomial con probabilidad dada por la media condicional $\pi(x)$.

1.2. Ajuste del modelo de regresión logística

Supongamos que tenemos una muestra de n observaciones independientes de la pareja (x_i, y_i), $i = 1, 2, ..., n$, donde y_i denota el valor de la variable de respuesta que puede ser cero o uno y x_i es el valor de la variable independiente para la i-ésima observación. Para ajustar el modelo de regresión logística en la ecuación 1.1 a un conjunto de datos, se requiere que estimemos los valores de β_0 y β_1, los parámetros desconocidos en el modelo.

En regresión lineal el método más utilizado para estimar los parámetros

desconocidos es el de **mínimos cuadrados**. En este método se eligen los valores de β_0 y β_1 que minimizan la suma de los cuadrados de las diferencias entre los valores observados de Y y los valores esperados basados en el modelo. Bajo las suposiciones usuales del modelo de regresión lineal, el método de mínimos cuadrados produce estimadores con propiedades estadísticas deseables. Desafortunadamente, cuando el método de mínimos cuadrados es aplicado a un modelo con variable de respuesta dicotómica, los estimadores obtenidos no tienen estas mismas propiedades.

El método general de estimación que conduce a la función de mínimos cuadrados en el modelo de regresión lineal (bajo el supuesto de que los errores siguen una distribución normal) es llamado **máxima verosimilitud**. En un sentido muy general, el método de máxima verosimilitud produce valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto observado de datos. Para poder aplicar este método, primero debemos construir una función llamada **función de verosimilitud**. Esta función expresa la probabilidad de obtener los datos observados como función de los parámetros desconocidos. Los **estimadores de máxima verosimilitud** de estos parámetros son aquellos que maximizan la función de verosimilitud. Por lo tanto, los estimadores resultantes serán aquellos que concuerdan de manera más precisa con los datos observados. Ahora describiremos como encontrar estos valores para el modelo de regresión logística.

Si Y toma los valores cero o uno entonces la expresión para $\pi(x)$ dada en la ecuación 1.1 proporciona (para un valor arbitrario de $\boldsymbol{\beta}' = (\beta_0, \beta_1)$, el vector de parámetros) la probabilidad condicional de que Y sea igual a 1 dado x. Esto se denotará como $P(Y = 1 \mid x)$. De aquí se sigue que la cantidad $1 - \pi(x)$ es la probabilidad condicional de que Y sea igual a cero dado x, $P(Y = 0 \mid x)$. Por lo tanto, para aquellas parejas (x_i, y_i), donde $y_i = 1$, la contribución a la función de verosimilitud es $\pi(x_i)$, y para aquellas parejas donde $y_i = 0$ la contribución a la función de verosimilitud es $1 - \pi(x_i)$, donde la cantidad $\pi(x_i)$ denota el valor de $\pi(x)$ evaluado en x_i. Una forma

conveniente de expresar la contribución a la función de verosimilitud de las parejas (x_i, y_i) es a través del término

$$\zeta(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{1.2}$$

Si suponemos que las observaciones son independientes, la función de verosimilitud es obtenida del producto de los términos dados en la expresión 1.2 como sigue:

$$l(\beta) = \prod_{i=1}^{n} \zeta(x_i) \tag{1.3}$$

El principio de máxima verosimilitud establece que el estimador de $\boldsymbol{\beta}$ es aquel valor que maximiza la expresión de la ecuación 1.3. Sin embargo, matemáticamente es más fácil trabajar con el logaritmo de la ecuación 1.3. Esta expresión, la **log verosimilitud**, es la siguiente:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{1.4}$$

Para encontrar el valor de $\boldsymbol{\beta}$ que maximiza el valor $L(\boldsymbol{\beta})$, derivamos $L(\boldsymbol{\beta})$ con respecto a β_o y β_1 e igualamos a cero la expresión resultante (por el criterio de la primera derivada[1]), es decir:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \frac{\partial L(\boldsymbol{\beta})}{\partial \pi(x_i)} \cdot \frac{\partial \pi(x_i)}{\partial \beta_0}$$

donde

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \pi(x_i)} = \sum_{i=1}^{n} \left\{ \frac{y_i - \pi(x_i)}{\pi(x_i) - [1 - \pi(x_i)]} \right\}$$

y

$$\frac{\partial \pi(x_i)}{\partial \beta_0} = \pi(x_i)[1 - \pi(x_i)]$$

[1]El criterio de la primera derivada establece lo siguiente: si $f(x)$ es una función diferenciable en (a, b) y c es un máximo relativo de (a, b), entonces, si $f'(c)$ existe, $f'(c) = 0$

entonces

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} \left\{ \frac{y_i - \pi(x_i)}{\pi(x_i) - [1 - \pi(x_i)]} \right\} \cdot \pi(x_i)[1 - \pi(x_i)] = \sum_{i=1}^{n} [y_i - \pi(x_i)]$$

De esta forma

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{1.5}$$

Por otro lado

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} = \frac{\partial L(\boldsymbol{\beta})}{\partial \pi(x_i)} \cdot \frac{\partial \pi(x_i)}{\partial \beta_1}$$

donde

$$\frac{\partial \pi(x_i)}{\partial \beta_1} = x_i \pi(x_i)[1 - \pi(x_i)]$$

entonces

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} \left\{ \frac{y_i - \pi(x_i)}{\pi(x_i) - [1 - \pi(x_i)]} \right\} \cdot x_i \pi(x_i)[1 - \pi(x_i)] = \sum_{i=1}^{n} x_i [y_i - \pi(x_i)]$$

Así

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \tag{1.6}$$

Las ecuaciones 1.5 y 1.6 son llamadas ecuaciones de verosimilitud.

En regresión lineal, las **ecuaciones de verosimilitud**, obtenidas al derivar la suma de las desviaciones al cuadrado con respecto a β, son lineales en los parámetros desconocidos y por lo tanto fáciles de resolver. Para el caso de regresión logística las expresiones en las ecuaciones 1.5 y 1.6 son no lineales en β_0 y β_1 y por lo tanto requieren de un algoritmo especial para la solución.

El valor de $\boldsymbol{\beta}$ obtenido en la solución de las ecuaciones 1.5 y 1.6 es conocido como el estimador de máxima verosimilitud y se denotará como $\widehat{\boldsymbol{\beta}}$. En general, el uso del símbolo $\widehat{\ }$ denotará el estimador de máxima verosimilitud de la respectiva cantidad. Por ejemplo, $\widehat{\pi}(x_i)$ es el estimador de máxima

verosimilitud de $\pi(x_i)$. Esta cantidad proporciona una estimación de la probabilidad condicional de que Y sea igual a 1, dado que x es igual a x_i. Una consecuencia interesante de la ecuación 1.5 es que

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\pi}(x_i)$$

Esto es, la suma de los valores observados de y es igual a la suma de sus valores esperados.

1.2.1. Algoritmo para obtener los estimadores de los parámetros del modelo de regresión logística

Supongamos que se van a usar n observaciones para estimar los valores de q parámetros, $\beta_{1,}\beta_2, ..., \beta_q$, y denotemos el log de verosimilitud por $L(\boldsymbol{\beta})$. Las q derivadas de la función de log verosimilitud con respecto a $\beta_{1,}\beta_2, ..., \beta_q$, son llamadas los **scores eficientes**, y pueden ser agrupados para formar el vector de $q \times 1$ de scores eficientes, cuyo j-ésimo componente es $\partial L(\boldsymbol{\beta}) / \partial \beta_j$, para $j = 1, 2, ..., q$. Denotemos este vector por $\mathbf{u}(\boldsymbol{\beta})$. Ahora, sea $\mathbf{H}(\boldsymbol{\beta})$ la matriz de $q \times q$ de segundas derivadas parciales de $L(\boldsymbol{\beta})$, donde el (j,k)-ésimo elemento de $\mathbf{H}(\boldsymbol{\beta})$ es

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_u}$$

para $j = 1, 2, ...q$; $k = 1, 2, ...q$. La matriz $\mathbf{H}(\boldsymbol{\beta})$ también es conocida como **matriz Hessiana.**

Consideremos $\mathbf{u}\left(\widehat{\boldsymbol{\beta}}\right)$, el vector de scores eficientes evaluado en el estimador de máxima verosimilitud de $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$. Usando series de Taylor para expandir $\mathbf{u}\left(\widehat{\boldsymbol{\beta}}\right)$ alrededor de $\boldsymbol{\beta}^*$, donde $\boldsymbol{\beta}^*$ es un valor cercano a $\widehat{\boldsymbol{\beta}}$, obtenemos:

$$\mathbf{u}\left(\widehat{\boldsymbol{\beta}}\right) \approx \mathbf{u}(\boldsymbol{\beta}^*) + \mathbf{H}(\boldsymbol{\beta}^*)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \tag{1.7}$$

Por definición, los estimadores de máxima verosimilitud de las $\beta^{'s}$ deben

satisfacer las ecuaciones

$$\frac{\partial L\left(\boldsymbol{\beta}\right)}{\partial \beta_j} \mid_{\widehat{\beta}} = 0$$

para $j = 1, 2, ..., q$, y así, $\mathbf{u}\left(\widehat{\boldsymbol{\beta}}\right) = 0$. De la ecuación 1.7, se sigue entonces que

$$\widehat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}^* - \mathbf{H}^{-1}\left(\boldsymbol{\beta}^*\right)\mathbf{u}(\boldsymbol{\beta}^*)$$

lo cual sugiere un esquema iterativo para la estimación de $\widehat{\boldsymbol{\beta}}$, en el cual el estimador de $\boldsymbol{\beta}$ en el $(r+1)$-ésimo ciclo de la iteración está dado por

$$\widehat{\boldsymbol{\beta}}_{r+1} = \widehat{\boldsymbol{\beta}}_r - \mathbf{H}^{-1}\left(\widehat{\boldsymbol{\beta}}_r\right)\mathbf{u}(\widehat{\boldsymbol{\beta}}_r) \tag{1.8}$$

para $r = 0, 1, 2, ...$, donde $\widehat{\boldsymbol{\beta}}_0$ es un vector de estimadores iniciales de $\boldsymbol{\beta}$. Éste es el **procedimiento de Newton-Raphson** para obtener el estimador de máxima verosimilitud de $\boldsymbol{\beta}$.

El algoritmo usado por la mayoría de los paquetes de cómputo estadístico para el ajuste del modelo de regresión logística, es una modificación de este esquema, en el cual $\mathbf{H}\left(\boldsymbol{\beta}\right)$ es reemplazada por la matriz de valores esperados de las segundas derivadas parciales de la función de log verosimilitud. Cuando esta matriz es multiplicada por -1, obtenemos la **matriz de información,** cuyo (j,k)-ésimo elemento es de la forma:

$$-E\left\{\frac{\partial^2 \log \mathbf{L}\left(\boldsymbol{\beta}\right)}{\partial \beta_j \partial \beta_k}\right\}$$

para $j = 1, 2, ...q$. Esta matriz se denota por $\mathbf{I}\left(\boldsymbol{\beta}\right)$, y juega un papel importante en la estimación por máxima verosimilitud, puesto que la inversa de $\mathbf{I}\left(\boldsymbol{\beta}\right)$ es la **matriz de varianzas y covarianzas asintótica** de los estimadores de máxima verosimilitud de los parámetros. Usando la matriz de información en el esquema iterativo definido en la ecuación 1.8, se obtiene

$$\widehat{\boldsymbol{\beta}}_{r+1} = \widehat{\boldsymbol{\beta}}_r + \mathbf{I}^{-1}\left(\widehat{\boldsymbol{\beta}}_r\right)\mathbf{u}(\widehat{\boldsymbol{\beta}}_r) \tag{1.9}$$

El procedimiento iterativo basado en la ecuación 1.9 es conocido como el **método de scores de Fisher.** Notemos que como una consecuencia de este

esquema iterativo, podemos obtener un estimador de la matriz de varianzas y covarianzas asintótica, $\mathbf{I}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)$.

Los dos métodos, el de Newton-Raphson y el de scores de Fisher, convergen al estimador de máxima verosimilitud de $\boldsymbol{\beta}$.

1.3. Prueba de significancia de los coeficientes

Una vez estimados los coeficientes del modelo, nos interesa evaluar la significancia de las variables independientes en el modelo ajustado. Esto involucra la formulación y prueba de hipótesis estadísticas para determinar si las variables independientes del modelo están "significativamente" relacionadas a la variable de respuesta.

Una pregunta que nos lleva a probar la significancia de las variables en cualquier modelo es la siguiente: *¿el modelo que incluye la variable en cuestión nos dice más acerca de la variable de respuesta que aquel modelo que no la incluye?* Esta pregunta puede ser contestada comparando los valores observados de la variable de respuesta con los valores ajustados por cada uno de los dos modelos: el primero incluyendo la variable en cuestión y el segundo sin ella. Si los valores ajustados con la variable en el modelo son mejores, o más precisos en algún sentido que cuando la variable no está incluida en el modelo, entonces decimos que la variable en cuestión es "significativa". Es importante hacer notar que no estamos considerando si los valores ajustados son una representación precisa de los datos observados en un sentido absoluto (esto es llamado **bondad de ajuste**). Nuestra pregunta está planteada en un sentido relativo.

El método general para probar la significancia de variables es fácilmente ilustrado en el modelo de regresión lineal, y su uso motivará el enfoque usado para la regresión logística. Una comparación de los dos enfoques resaltará las diferencias entre modelar una variable de respuesta continua y una dicotómica.

En regresión lineal la evaluación de la significancia de las variables se hace a través de lo que se conoce como **tabla de análisis de varianza.** Esta tabla particiona la suma de cuadrados total de las desviaciones de las observaciones alrededor de la media en dos partes: (1) la suma de las desviaciones al cuadrado de las observaciones alrededor de la línea de regresión (o suma de los cuadrados de los residuales que denotaremos por SCE), y (2) la suma de los cuadrados de los valores ajustados, basados en el modelo de regresión lineal, alrededor de la media de la variable dependiente (o suma de cuadrados debida a la regresión que denotaremos por SCR). Ésta es una forma conveniente de exhibir la comparación entre valores observados y valores ajustados bajo los dos modelos. En regresión lineal la comparación de valores observados contra ajustados está basada en el cuadrado de la distancia entre los dos. Si y_i denota el valor observado y $\widehat{y_i}$ denota el valor ajustado para la i-ésima observación, entonces la estadística usada para evaluar esta comparación es:

$$SCE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

En el modelo que no contiene a la variable dependiente en cuestión el único parámetro es β_0, y $\widehat{\beta_0} = \overline{y}$, la media de la variable de respuesta. En este caso $\widehat{y_i} = \overline{y}$ y SCE es igual a la varianza total. Cuando incluimos a la variable independiente en el modelo, cualquier decremento en SCE será debido al hecho de que la pendiente del coeficiente de la variable independiente es diferente de cero. El cambio en el valor de SCE es debido a la fuente de variabilidad de la regresión, denotado por SCR. Esto es:

$$SCR = \left[\sum_{i=1}^{n} (y_i - \overline{y})^2\right] - \left[\sum_{i=1}^{n} (y_i - \widehat{y_i})^2\right]$$

En regresión lineal, el interés se centra en el tamaño de SCR. Un valor grande sugiere que la variable independiente es importante, mientras que un valor pequeño sugiere que la variable independiente no es útil para predecir a la variable de respuesta.

El principio usado en regresión logística es el mismo: *comparar los valores observados de la variable de respuesta con los valores ajustados obtenidos de los modelos con la variable en cuestión y sin ella.* En regresión logística la comparación de valores observados contra valores ajustados está basada en la función de log verosimilitud definida en la ecuación 1.4. Para entender mejor esta comparación, pensemos que un valor observado de la variable de respuesta es también un valor ajustado que resulta de un **modelo saturado.** Un modelo saturado es aquel que contiene tantos parámetros como datos existen. (Un ejemplo de modelo saturado es ajustar un modelo de regresión lineal simple cuando sólo existen dos datos disponibles, $n = 2$)

La comparación de valores observados contra ajustados usando la función de verosimilitud está basada en la siguiente expresión:

$$D = -2 \ln \left[\frac{\text{verosimilitud del modelo actual}}{\text{verosimilitud del modelo saturado}} \right] \tag{1.10}$$

La cantidad dentro de los corchetes en la expresión anterior es llamada **cociente de verosimilitudes.** Usar menos dos veces su logaritmo es necesario para obtener una cantidad cuya distribución sea conocida y por lo tanto pueda ser usada con el propósito de probar hipótesis. Tal prueba es llamada la **prueba del cociente de verosimilitudes.** Usando la ecuación 1.4, la ecuación 1.10 queda como sigue:

$$D = -2 \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\widehat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \widehat{\pi}_i}{1 - y_i} \right) \right] \tag{1.11}$$

donde $\widehat{\pi}_i = \widehat{\pi}(x_i)$.

La estadística D, en la ecuación 1.11 es llamada por algunos autores **la devianza** y juega un papel muy importante en la evaluación de la bondad de ajuste. La devianza, para la regresión logística juega el mismo papel que la suma de los cuadrados de residuales en regresión lineal. De hecho cuando la desviación es evaluada para la regresión lineal, se tiene que es idénticamente igual a SCE.

Para propósitos de evaluar la significancia de una variable independiente, comparamos el valor de D con y sin la variable independiente en la ecuación. El cambio en D debido a la inclusión de la variable independiente en el modelo se obtiene como sigue:

$$G = D\,(\text{para el modelo sin la variable}) - D\,(\text{para el modelo con la variable})$$

Esta estadística tiene el mismo papel en regresión logística que el numerador de la estadística F para regresión lineal. Como la verosimilitud del modelo saturado es común para ambos valores de D, G puede ser expresada como:

$$G = -2\ln\left[\frac{\text{verosimilitud sin la variable}}{\text{verosimilitud con la variable}}\right] \tag{1.12}$$

Para el caso específico de una sola variable independiente, es fácil mostrar que cuando esta variable no se incluye en el modelo, el estimador de máxima verosimilitud para β_0 es $\ln(n_1/n_0)$, donde $n_1 = \sum y_i$ y $n_0 = \sum (1 - y_i)$ y que el valor ajustado es constante, n_1/n. En este caso el valor de G es:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi}_i^{y_i}\left(1-\widehat{\pi}_i\right)^{(1-y_i)}}\right] \tag{1.13}$$

o

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i\ln\left(\widehat{\pi}_i\right) + (1-y_i)\ln\left(1-\widehat{\pi}_i\right)\right] - \left[n_1\ln\left(n_1\right) + n_0\ln\left(n_0\right) - n\ln\left(n\right)\right]\right\} \tag{1.14}$$

Bajo la hipótesis de que β_1 es igual a cero, la estadística G tendrá una distribución Ji-cuadrada con 1 grado de libertad (suponiendo que se tiene una muestra suficientemente grande). Utilizaremos el símbolo $\chi^2(v)$, para denotar una distribución Ji-cuadrada con v grados de libertad. Con esta notación, la regla de decisión para la prueba de hipótesis: $H_0 : \beta_1 = 0\ vs\ H_1 : \beta_1 \neq 0$ será:

Rechazar H_0 a un nivel de significancia α si:

$$G > \chi^2_{\alpha/2}(1)$$

donde G es la estadística definida en la ecuación 1.14, $\chi^2_{\alpha/2}(v)$ es el cuantil $\alpha/2$ de una $\chi^2(1)$ y $\chi^2_{1-\alpha/2}(1)$ es el cuantil $1-\alpha/2$ de una $\chi\ ^2(1)$.

Con lo anterior, sólo se muestra la evidencia estadística de que la variable es significativa, pero para incluirla en el modelo también se deben tomar en cuenta otros factores importantes como el ajuste del modelo y la inclusión de otras variables potencialmente importantes.

Otras dos pruebas estadísticamente similares han sido sugeridas: la prueba de Wald y la prueba de Score. Las suposiciones que se hacen en estas dos pruebas son las mismas que se hacen para la prueba del cociente de verosimilitudes en 1.13. Una breve descripción de estas pruebas se presenta a continuación.

La prueba de Wald se obtiene comparando el estimador de máxima verosimilitud de la pendiente del parámetro, $\widehat{\beta}_1$, con un estimador de su error estándar. El cociente resultante, bajo la hipótesis de que $\beta_1 = 0$, tiene una distribución normal estándar. En este trabajo no se ha discutido la forma de obtener el estimador de los errores estándar de los parámetros estimados, pero pueden evaluarse fácilmente en la mayoría de los paquetes de software estadístico. Así la estadística de la prueba de Wald para el modelo de regresión logística es:

$$W = \frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)}$$

y el p-value es $P(|Z| > W)$, donde Z denota una variable aleatoria con distribución normal estándar. Algunos autores que han evaluado el desempeño de la prueba de Wald, han encontrado que su comportamiento es muy cuestionable, pues a menudo cae en el error de no detectar la significancia del coeficiente cuando en realidad si lo era, por lo tanto recomiendan usar la prueba del cociente de verosimilitudes.

Tanto la prueba del cociente de verosimilitudes como la prueba de Wald requieren del estimador de máxima verosimilitud para β_1. Para el caso de una sola variable, esto no es difícil o computacionalmente costoso, sin embargo para un conjunto de datos grande con muchas variables, el número de

iteraciones para obtener los coeficientes de máxima verosimilitud puede ser considerable.

Una prueba para la significancia de variables que no requiere de estos cálculos es la prueba de Score. Los que proponen esta prueba dicen que la reducción de cálculos es su mayor ventaja. Sin embargo, el uso de esta prueba está limitado por el hecho de que no puede ser obtenida fácilmente de algunos paquetes de software. La prueba de Score está basada en la teoría de distribución de las derivadas del log verosimilitud. En general, ésta es una prueba que requiere del cálculo de matrices.

En el caso univariado, esta prueba está basada en la distribución condicional de la derivada en la ecuación 1.6, dada la derivada en la ecuación 1.5. La prueba usa el valor de la ecuación 1.6, evaluado al usar $\beta_0 = \ln(n_1/n_0)$ y $\beta_1 = 0$. Como se observó antes, bajo estos valores de los parámetros, $\widehat{\pi} = n_1/n = \overline{y}$. Por lo tanto, el lado izquierdo de la ecuación 1.6 se convierte en $\sum x_i (y_i - \overline{y})$. Se puede demostrar que la varianza estimada es $\overline{y}(1-\overline{y}) \sum (x_i - \overline{x})^2$. Así pues, la estadística para la prueba de Score (S) es:

$$S = \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y}) \sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

y su p-value es $P(|Z| > S)$ donde Z es una variable aleatoria con distribución normal estándar.

En resumen, el procedimiento para probar la significancia del coeficiente de una variable en regresión logística es similar al usado en regresión lineal, solo que usa la función de verosimilitud para una variable de respuesta dicotómica.

1.4. El modelo múltiple de regresión logística

En las secciones anteriores hemos introducido el modelo de regresión logística en el contexto univariado. Como en el caso de regresión lineal, la mayor

virtud de esta técnica está en la capacidad de modelar varias variables, algunas de las cuales pueden estar en diferentes escalas de medida. En esta sección generalizaremos el modelo de regresión logística al caso de más de una variable independiente. Éste es conocido como el "caso múltiple". Una consideración importante del modelo logístico múltiple será la estimación de los coeficientes del modelo y probar su significancia. Esto se hará siguiendo el mismo razonamiento que el modelo univariado. Una consideración adicional que se hará en esta sección será el uso de variables de diseño para modelar variables independientes nominales. En todos los casos se va a suponer que existe una colección de variables que será examinada.

Consideremos una colección de p variables independientes las cuales serán denotadas por el vector $\mathbf{x}' = (x_1, x_2, ..., x_p)$. Por el momento supondremos que cada una de estas variables está definida en un intervalo. Denotemos la probabilidad de que la variable de respuesta sea uno, dado $\mathbf{x}$, como $P(Y = 1 \mid \mathbf{x}) = \pi(\mathbf{x})$. Entonces el logit del modelo múltiple de regresión logística está dado por la ecuación:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{1.15}$$

y para este caso

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{1.16}$$

Si algunas de las variables independientes son variables nominales tales como sexo, grupo de tratamiento, nivel de estudios, etc., entonces sería inapropiado incluirlas en el modelo, pues los números utilizados para representarlas son solamente identificadores y no tienen ningún significado numérico. En esta situación lo que se hace es utilizar un conjunto de **variables mudas** (o **variables dummies**). Por ejemplo, supongamos que una de las variables independientes es el nivel de estudios, el cual ha sido denotado como "básico", "medio superior" o "superior". En este caso se necesitarán dos variables *dummy*. Una posible estrategia de asignación es que cuando en nivel se estudios sea "básico", las dos variables dummy, D_1 y D_2 sean cero; cuando el

nivel de estudios sea "medio superior", D_1 sea igual a 1 y D_2 sea igual a cero y cuando el nivel de estudios sea "superior" se utilice $D_1 = 0$ y $D_2 = 1$. La mayor parte del software para regresión logística genera las variables dummy, y otros tienen la opción para utilizar métodos diferentes.

En general, si una variable nominal tiene k posibles valores, entonces se necesitarán $k - 1$ variables dummy. La notación para indicar el uso de variables dummy es la siguiente: supongamos que la j-ésima variable independiente, x_j, puede tomar k_j posibles valores. Entonces las $k_j - 1$ variables dummy serán denotadas por D_{ju} y los coeficientes de estas variables dummy serán denotados por β_{ju}, $u = 1, 2, ..., k_j - 1$. Por lo tanto, el logit para el modelo con p variables y la j-ésima variable nominal, sería:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \sum_{u=1}^{k_j - 1} \beta_{ju} D_{ju} + \beta_p x_p$$

Cuando discutimos el modelo múltiple de regresión logística, en general y para fines prácticos, suprimimos la suma que indica el uso de variables *dummy*. La excepción a esto será la discusión de estrategias de modelación cuando se necesite usar valores específicos de los coeficientes para cualquier variable *dummy* en el modelo.

1.4.1. Ajuste del modelo múltiple de regresión logística

Supongamos que tenemos una muestra de n observaciones independientes de la pareja $(\mathbf{x}_i, y_i)$, $i = 1, 2, ..., n$. Como en el caso univariado, ajustar el modelo requiere que obtengamos los estimadores del vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_p)$. El método de estimación usado en el caso múltiple será el mismo que en el caso univariado, máxima verosimilitud. La función de verosimilitud es casi idéntica a la de la ecuación 1.3, con la diferencia de que ahora $\pi(\mathbf{x})$ está definido como en la ecuación 1.16. Tendremos $p + 1$ ecuaciones de verosimilitud que son obtenidas al derivar la función de log verosimilitud con respecto a los $p + 1$ coeficientes. Las ecuaciones de verosimilitud que resultan pueden ser

expresadas como sigue:

$$\sum_{i=1}^{n} \left[y_i - \pi \left(\mathbf{x}_i \right) \right] = 0$$

y

$$\sum_{i=1}^{n} x_{ij} \left[y_i - \pi \left(\mathbf{x}_i \right) \right] = 0$$

para $j = 1, 2, .., p$.

Como en el modelo univariado, las ecuaciones de verosimilitud se pueden resolver utilizando el método de Newton-Raphson. Sea $\widehat{\boldsymbol{\beta}}$ la solución de estas ecuaciones. Entonces los valores ajustados para el modelo múltiple de regresión logística son $\widehat{\pi}\left(x_i \right)$, el valor de la ecuación 1.16 evaluada usando $\widehat{\boldsymbol{\beta}}$ y x_i.

Hasta este punto sólo se ha hecho una breve mención del método utilizado para estimar los errores estándar de los coeficientes estimados. Ahora que el modelo de regresión logística ha sido generalizado en concepto y notación al caso múltiple, consideraremos la estimación de errores estándar con más detalle.

El método de estimación de varianzas y covarianzas de los coeficientes estimados está basado en la teoría de estimación por máxima verosimilitud. Esta teoría dice que los estimadores se obtienen de la matriz de segundas derivadas parciales de la función de log verosimilitud. Estas derivadas parciales tienen la siguiente forma general:

$$\frac{\partial^2 L \left(\boldsymbol{\beta} \right)}{\partial \beta_j^2} = - \sum_{i=1}^{n} x_{ij}^2 \pi_i \left(1 - \pi_i \right) \tag{1.17}$$

y

$$\frac{\partial^2 L \left(\boldsymbol{\beta} \right)}{\partial \beta_j \partial \beta_u} = - \sum_{i=1}^{n} x_{ij} x_{iu} \pi_i \left(1 - \pi_i \right) \tag{1.18}$$

para $j, u = 0, 1, 2, ..., p$ donde $\pi_i = \pi \left(x_i \right)$. Denotemos como $\mathbf{I} \left(\boldsymbol{\beta} \right)$ a la matriz de $(p+1)$ por $(p+1)$ que contiene los negativos de los términos dados en las ecuaciones 1.17 y 1.18. Esta matriz es conocida como la **matriz de información**. Las varianzas y covarianzas de los coeficientes estimados se obtienen de

la inversa de esta matriz, la cual se denota como $\Sigma(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. Excepto en casos muy especiales, no es posible escribir explícitamente una expresión para los elementos de esta matriz. Por lo tanto, utilizaremos la notación $\sigma^2\left(\beta_j\right)$ para denotar el j-ésimo elemento diagonalde $\Sigma(\boldsymbol{\beta})$, el cual es la varianza de $\widehat{\beta}_j$, y $\sigma\left(\beta_j, \beta_u\right)$ para denotar un elemento arbitrario fuera de la diagonal, es cual es la covarianza entre $\widehat{\beta}_j$ y $\widehat{\beta}_u$. Los estimadores de las varianzas y covarianzas, los cuales serán denotados por $\widehat{\Sigma}\left(\widehat{\boldsymbol{\beta}}\right)$, son obtenidos evaluando $\Sigma(\boldsymbol{\beta})$ en $\widehat{\boldsymbol{\beta}}$. Usaremos $\widehat{\sigma}^2\left(\widehat{\beta}_j\right)$ y $\widehat{\sigma}^2\left(\widehat{\beta}_j, \widehat{\beta}_u\right), j, u = 0, 1, 2, ..., p$, para denotar los valores de $\widehat{\Sigma}\left(\widehat{\boldsymbol{\beta}}\right)$. Entonces, los estimadores de los errores estándar de los coeficientes estimados están denotados por la siguiente expresión:

$$\widehat{SE}\left(\widehat{\beta}_j\right) = \left[\widehat{\sigma}^2\left(\widehat{\beta}_j\right)\right]^{1/2} \tag{1.19}$$

para $j = 0, 1, 2, ..., p$.

Otra manera de escribir la matriz de información y que será útil cuando se discuta el ajuste del modelo es $\widehat{\mathbf{I}}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{X}'\mathbf{V}\mathbf{X}$ donde $\mathbf{X}$ es una matriz de n por $p+1$ que contiene los datos de cada unidad observada, y $\mathbf{V}$ es una matriz diagonal de n x n que contiene los elementos $\widehat{\pi}_i(1-\widehat{\pi}_i)$. Esto es:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{21} & \cdots & x_{2p} \\ & & \ddots & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

y

$$\mathbf{V} = \begin{bmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{bmatrix}$$

1.4.2. Prueba de significancia del modelo

Una vez que hemos ajustado un modelo múltiple de regresión logística, empezamos el proceso de evaluación del mismo. Como en el caso univariado,

el primer paso en este proceso es usualmente la evaluación de la significancia de las variables en el modelo. La prueba del cociente de verosimilitudes para la significancia global de los p coeficientes de las variables independientes del modelo se lleva a cabo exactamente de la misma manera que en el caso univariado. La prueba está basada en la estadística G dada en las ecuaciones 1.11 y 1.12. La única diferencia es que los valores ajustados, $\widehat{\pi}$, bajo el modelo, están basados en un vector que contiene $p+1$ parámetros, $\widehat{\boldsymbol{\beta}}$. Bajo la hipótesis de que los p coeficientes de las covariables en el modelo son iguales a cero, la distribución de G será Ji-cuadrada con p grados de libertad. Rechazar la hipótesis nula en este caso, tiene una interpretación análoga a la de regresión lineal múltiple; podemos concluir que por lo menos uno, o tal vez todos los coeficientes de las variables independientes son diferentes de cero.

Antes de concluir que alguno o todos los coeficientes son diferentes de cero, sería bueno ver la estadística univariada de Wald, $W_j = \widehat{\beta}_j / \widehat{ES}\left(\widehat{\beta}_j\right)$. Bajo la hipótesis nula de que un coeficiente es igual a cero, esta estadística sigue una distribución normal estándar. Por lo tanto, el valor de esta estadística puede darnos una indicación de cuales de las variables en el modelo pueden o no ser significativas.

Considerando que nuestra meta es obtener el mejor ajuste en el modelo mientras minimizamos el número de parámetros, el siguiente paso lógico sería ajustar un modelo reducido que contenga sólo aquellas variables que resultaron ser significativas y compararlo con el modelo que contiene todas las variables.

Siempre que una variable nominal independiente sea incluida (o excluida) de un modelo, todas sus variables *dummy* deberán ser incluidas (o excluidas), no hacerlo implica que le hemos reasignado nuevos valores a esta variable. Por ejemplo, si la variable nominal es raza y sus variables *dummy* son $D_1 =$ "$Blanco$", $D_2 =$ "$Negro$", $D_3 =$ "$Otro$", y sólo incluimos D_1, entonces lo que estamos diciendo es que la variable nominal es interpretada como "$Blanco$" o "$No\ Blanco$". Si el número de variables *dummy* de la variable nominal es k,

entonces la contribución a los grados de libertad en la prueba del cociente de verosimilitudes por la exclusión de esta variable será $k-1$. De esta manera, si excluyéramos la variable raza de nuestro modelo, habría dos grados de libertad para la prueba.

Anteriormente describimos, para el caso univariado, una prueba equivalente a la prueba del cociente de verosimilitudes para evaluar la significancia del modelo; la prueba de Wald. Ahora discutiremos brevemente la versión múltiple de esta prueba.

El caso multivariado análogo de la prueba de Wald se obtiene del siguiente cálculo vector-matricial

$$\begin{aligned} W &= \widehat{\boldsymbol{\beta}}' \left[\widehat{\Sigma}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1} \widehat{\boldsymbol{\beta}} \\ &= \widehat{\boldsymbol{\beta}}' \left(\mathbf{X}'\mathbf{V}\mathbf{X}\right) \widehat{\boldsymbol{\beta}} \end{aligned}$$

La estadística W se distribuye como una Ji-cuadrada con $p+1$ grados de libertad bajo la hipótesis de que cada uno de los $p+1$ coeficientes es igual a cero. La prueba para solamente los p coeficientes de las variables independientes se obtiene eliminando $\widehat{\beta}_0$ de $\widehat{\beta}$ y el primer renglón y primera columna de $(\mathbf{X}'\mathbf{V}\mathbf{X})$. Puesto que la evaluación de esta prueba requiere la capacidad de llevar a cabo operaciones vector-matriciales y de obtener $\widehat{\beta}$, no tiene ninguna ventaja sobre la prueba del cociente de verosimilitudes.

Capítulo 2

Discriminación y clasificación

2.1. Introducción

Supongamos que un banco está interesado en determinar si un cliente es sujeto de crédito. El banco puede saber algunas cosas acerca del cliente como su ingreso, edad, número de tarjetas de crédito con las que cuenta y el tamaño de su familia. Con base en estos criterios, al banco le gustaría saber si 1) ¿Se puede usar esta información para construir una regla que clasificará a nuevos clientes como "sujetos de crédito" o "no sujetos de crédito"? 2) ¿Cuál es la regla para clasificar nuevos clientes? 3) ¿Cuáles son las posibilidades de cometer equivocaciones al aplicar la regla?

Notemos que, en el ejemplo anterior, se cometen equivocaciones siempre que se clasifique un nuevo cliente en la población errónea, esto es, se comete un error cuando se predice que un cliente es sujeto de crédito cuando en realidad no lo es, o bien, cuando un cliente que si es sujeto de crédito se predice como no sujeto de crédito. Observemos también que es probable que estos dos tipos de errores no sean igual de graves. Puede no ser muy grave negarle el crédito a una persona que en realidad si es sujeta de éste, pero podría ser demasiado grave otorgarle crédito a una persona cuando en realidad no es sujeta de crédito, pues esto podría traerle pérdidas económicas al banco. El *análisis discriminante* es una técnica multivariada que se puede

usar para generar reglas con las que se pueda clasificar a lo clientes en la población apropiada.

En el análisis discriminante se desea poder predecir la pertenencia a una clase de una observación particular, con base en un conjunto de variables predictoras.

El análisis discriminante en ocasiones se conoce como *análisis de clasificación*. Supongamos que se tienen varias poblaciones de las que se pueden extraer observaciones. Supongamos también que se tiene una nueva observación que proviene de una de estas poblaciones, pero no se sabe de cuál. El objetivo básico del análisis discriminante es producir una regla o un esquema de clasificación que permita a un investigador predecir la población de la que es más probable que pertenezca la observación.

2.2. Discriminación para dos poblaciones.

Para fijar ideas, pensemos en una situación donde pudiéramos estar interesados en separar dos clases de objetos o asignar un nuevo objeto a una de las dos clases (o ambas). Es conveniente etiquetar las dos clases, digamos π_1 y π_2 . Los objetos son comúnmente separados o clasificados con base en algunas mediciones, por ejemplo, p variables aleatorias $\mathbf{X}' = [X_1, X_2, ...X_p]$. Los valores observados de $\mathbf{X}$ difieren en cierta medida de una clase a otra, pues si no fuera así podrían ser asignados a cualquier clase de manera indiscriminada. Podemos pensar en la totalidad de los valores de la primera clase como la población de valores $\mathbf{x}$ para π_1 y aquellos de la segunda clase como la población de valores $\mathbf{x}$ para π_2. Estas dos poblaciones pueden ser descritas por las funciones de densidad de probabilidad $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ y consecuentemente podemos hablar de asignar observaciones a poblaciones u objetos a clases alternativamente. Como un ejemplo de esta situación, supongamos que los clientes del banco antes mencionado son separados en dos poblaciones π_1: sujetos de crédito y π_2: no sujetos de crédito, con base en

valores observados de variables presumiblemente relevantes: ingreso, edad, número de tarjetas de crédito, tamaño de familia. En términos de *observación* y *población*, deseamos identificar una observación de la forma $\mathbf{x}' = [x_1(\text{ingreso}),\ x_2(\text{edad}),\ x_3(\text{tarjetas de crédito}),\ x_4(\text{tamaño de familia})]$ como elemento de la población π_1: sujeto de crédito o π_2: no sujeto de crédito.

La idea que tuvo Fisher fue transformar la observación multivariada $\mathbf{x}$ en una observación univariada y de tal manera que las y's derivadas de las poblaciones π_1 y π_2 estuvieran separadas tanto como fuera posible. Fisher propuso tomar combinaciones lineales de $\mathbf{x}$ para crear las y's, pues éstas serían simples funciones de $\mathbf{x}$ y por lo tanto matemáticamente fáciles de manejar. Al denotar por μ_{1Y} a la media de las Y's obtenidas de las X's que pertenecen a π_1 y por μ_{2Y} a la media de las Y's que pertenecen a π_2, Fisher seleccionó las combinaciones lineales que maximizaran el cuadrado de la distancia entre μ_{1Y} y μ_{2Y} relativa a la variabilidad de las Y's.

Empecemos definiendo

El valor esperado de una observación multivariada que proviene de π_1 como $\boldsymbol{\mu}_1 = E(\mathbf{X} \mid \pi_1)$

El valor esperado de una observación multivariada que proviene de π_2 como $\boldsymbol{\mu}_2 = E(\mathbf{X} \mid \pi_2)$

y supongamos que la matriz de varianzas y covarianzas

$$\Sigma = E(\mathbf{X} - \boldsymbol{\mu}_i)(\mathbf{X} - \boldsymbol{\mu}_i)',\ i = 1, 2 \tag{2.1}$$

es la misma para ambas poblaciones. Ahora consideremos la combinación lineal

$$\underset{(1\times 1)}{\mathbf{Y}} = \underset{(1\times p)(p\times 1)}{\boldsymbol{\lambda}'\ \mathbf{X}} \tag{2.2}$$

La media de Y es:

$$\mu_{1Y} = E(\mathbf{Y} \mid \pi_1) = E(\boldsymbol{\lambda}'\mathbf{X} \mid \pi_1) = \boldsymbol{\lambda}'\boldsymbol{\mu}_1$$

o

$$\mu_{2Y} = E(\mathbf{Y} \mid \pi_2) = E(\boldsymbol{\lambda}'\mathbf{X} \mid \pi_2) = \boldsymbol{\lambda}'\boldsymbol{\mu}_2$$

dependiendo de la población, pero su varianza

$$\sigma_Y^2 = Var\left(\boldsymbol{\lambda}'\mathbf{X}\right) = \boldsymbol{\lambda}' Var\left(\mathbf{X}\right)\boldsymbol{\lambda} = \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}$$

es la misma para ambas poblaciones.

La mejor combinación lineal se obtiene del cociente

$$\begin{aligned} \frac{\begin{array}{c}\text{Cuadrado de las distancias} \\ \text{entre las medias de } \mathbf{Y}\end{array}}{\text{Varianza de } \mathbf{Y}} &= \frac{\left(\mu_{1Y} - \mu_{2Y}\right)^2}{\sigma_Y^2} = \frac{\left(\boldsymbol{\lambda}'\boldsymbol{\mu}_1 - \boldsymbol{\lambda}'\boldsymbol{\mu}_2\right)^2}{\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}} \\ &= \frac{\boldsymbol{\lambda}'\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}} = \frac{\left(\boldsymbol{\lambda}'\boldsymbol{\delta}\right)^2}{\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}} \end{aligned} \quad (2.3)$$

donde $\boldsymbol{\delta} = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$ es la diferencia entre los vectores de medias. Notemos que la matriz $\boldsymbol{\delta}\boldsymbol{\delta}' = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'$ contiene los cuadrados y productos cruzados de las diferencias entre las medias de las poblaciones π_1 y π_2. Los coeficientes de la combinación lineal de Fisher

$\boldsymbol{\lambda}' = [\lambda_1, \lambda_2, ...\lambda_p]$ son tales que maximizan el cociente en 2.3, es decir, $\boldsymbol{\lambda}$ es de la forma:

$$\boldsymbol{\lambda} = c\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$$

para cualquier $c \neq 0$. Eligiendo $c = 1$, la combinación lineal resultante es:

$$\mathbf{Y} = \boldsymbol{\lambda}'\mathbf{X} = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}^{-1}\mathbf{X} \quad (2.4)$$

la cual es conocida como *función discriminante lineal de Fisher*.

La función discriminante lineal convierte a las poblaciones multivariadas π_1 y π_2 en poblaciones univariadas tales que sus medias estén separadas tanto como sea posible con respecto a la varianza de la población.

Definamos $y_0 = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0$ como el valor de la función discriminante para una nueva observación x_0 y

$$m = \frac{1}{2}\left(\mu_{1Y} - \mu_{2Y}\right) = \frac{1}{2}\left(\boldsymbol{\lambda}'\boldsymbol{\mu}_1 + \boldsymbol{\lambda}'\boldsymbol{\mu}_2\right) = \frac{1}{2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right) \quad (2.5)$$

el punto medio entre las medias de las dos poblaciones univariadas. Se puede demostrar que

$$E\left(\mathbf{Y}_0 \mid \pi_1\right) - m \geq 0$$

y

$$E\left(\mathbf{Y}_1 \mid \pi_2\right) - m < 0$$

Es decir, si $\mathbf{X}_0$ proviene de π_1, esperamos que $\mathbf{Y}_0$ sea mayor que el punto medio, y si $\mathbf{X}_0$ proviene de π_2, esperamos que $\mathbf{Y}_0$ sea menor que el punto medio. Por lo tanto, la regla de clasificación es:

$$\begin{aligned} \text{Asigne } \mathbf{x}_0 \text{ a } \pi_1 \text{ si } y_0 &= \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1} \mathbf{x}_0 \geq m \\ \text{Asigne } \mathbf{x}_0 \text{ a } \pi_2 \text{ si } y_0 &= \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1} \mathbf{x}_0 < m \end{aligned} \tag{2.6}$$

Alternativamente, podemos obtener m de y_0 y comparar el resultado con cero. En este caso, la regla se convierte en:

$$\begin{aligned} \text{Asigne } \mathbf{x}_0 \text{ a } \pi_1 \text{ si } y_0 - m &\geq 0 \\ \text{Asigne } \mathbf{x}_0 \text{ a } \pi_1 \text{ si } y_0 - m &< 0 \end{aligned} \tag{2.7}$$

En la práctica, las cantidades $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}$ rara vez son conocidas. Es por ello que las reglas en 2.6 y 2.7 no pueden ser llevadas a cabo a menos que λ y m puedan ser estimadas a partir de las observaciones que ya hemos clasificado correctamente.

Supongamos que tenemos n_1 observaciones de la variable aleatoria multivariada $\mathbf{X}' = [X_1, X_2, ...X_p]$ provenientes de la población π_1 y n_2 observaciones de la población π_2. Sus respectivas matrices de datos son:

$$\begin{aligned} \underset{(p\times n_1)}{\mathbf{X}_1} &= [\mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1n_1}] \\ \underset{(p\times n_2)}{\mathbf{X}_2} &= [\mathbf{x}_{21}, \mathbf{x}_{22}, ..., \mathbf{x}_{2n_2}] \end{aligned} \tag{2.8}$$

De estas matrices de datos, los vectores de medias muestrales y las matrices de varianzas y covarianzas muestrales están determinadas por:

$$\begin{aligned} \underset{(p\times 1)}{\overline{\mathbf{X}}_1} &= \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j} \qquad \underset{(p\times p)}{S_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} \left(\mathbf{x}_{1j} - \overline{\mathbf{x}}_1\right)\left(\mathbf{x}_{1j} - \overline{\mathbf{x}}_1\right)' \\ \underset{(p\times 1)}{\overline{\mathbf{X}}_2} &= \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j} \qquad \underset{(p\times p)}{S_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_1} \left(\mathbf{x}_{2j} - \overline{\mathbf{x}}_2\right)\left(\mathbf{x}_{2j} - \overline{\mathbf{x}}_2\right)' \end{aligned} \tag{2.9}$$

Puesto que suponemos que las poblaciones tienen la misma matriz de varianzas y covarianzas $\mathbf{\Sigma}$, las matrices de varianzas y covarianza muestrales S_1 y S_2 se combinan para derivar un sólo estimador insesgado de $\mathbf{\Sigma}$. En particular, el promedio ponderado

$$\mathbf{S}_p = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \tag{2.10}$$

$$= \frac{(n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2}{(n_1 + n_2 - 2)}$$

es un estimador insesgado de $\mathbf{\Sigma}$ si las matrices de datos $\mathbf{X}_1$ y $\mathbf{X}_2$ contienen muestras aleatorias de las poblaciones π_1 y π_2, respectivamente.

Las cantidades muestrales $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$ y $\mathbf{S}_p$ son substituidas por $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\mathbf{\Sigma}$ respectivamente en 2.4 para obtener así la *Función Lineal Discriminante Muestral de Fisher*

$$y = \widehat{\boldsymbol{\lambda}}\mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\,\mathbf{S}_p^{-1}\mathbf{x} \tag{2.11}$$

El punto medio, $\widehat{m}$, entre las medias muestrales univariadas, $\overline{y}_1 = \widehat{\boldsymbol{\lambda}}'\overline{\mathbf{x}}_1$ y $\overline{y}_2 = \widehat{\boldsymbol{\lambda}}'\overline{\mathbf{x}}_2$ está dado por

$$\widehat{m} = \frac{1}{2}\left(\overline{y}_1 + \overline{y}_2\right) = \frac{1}{2}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)'\mathbf{S}_p^{-1}\left(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2\right) \tag{2.12}$$

y la regla de clasificación basada en las muestras se convierte en

Asigne $\mathbf{x}_0$ a π_1 si

$$y_0 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\,\mathbf{S}_p^{-1}\mathbf{x}_0 \geq \widehat{m} = \frac{1}{2}\left(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\right)'\mathbf{S}_p^{-1}\left(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2\right)$$

o

$$y_0 - \widehat{m} \geq 0$$

Asigne $\mathbf{x}_0$ a π_2 si

$$y_0 < \widehat{m} \tag{2.13}$$

o

$$y_0 - \widehat{m} < 0$$

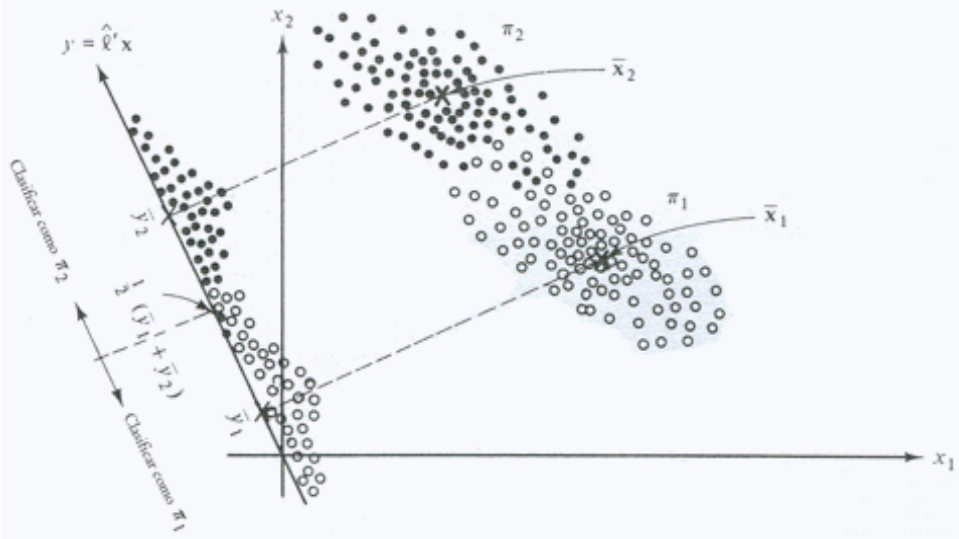

Tabla 2.1: Representación gráfica del procedimiento de Fisher para dos poblaciones con p=2

La solución que Fisher propuso a los problemas de separación y clasificación se ilustra en la figura 2.1 para p=2.

La función lineal discriminante muestral en 2.11 tiene la propiedad "óptima" de maximizar el cociente

$$\begin{aligned} \frac{\text{Cuadrado de la distancia entre las medias muestrales de } y}{\text{Varianza muestral de } y} &= \frac{(\overline{y}_1 - \overline{y}_2)^2}{s_y^2} \\ &= \frac{\left(\widehat{\boldsymbol{\lambda}}'\overline{\mathbf{x}}_1 - \widehat{\boldsymbol{\lambda}}'\overline{\mathbf{x}}_2\right)^2}{\widehat{\boldsymbol{\lambda}}'\mathbf{S}_p\widehat{\boldsymbol{\lambda}}} \\ &= \frac{\left(\widehat{\boldsymbol{\lambda}}'\mathbf{d}\right)^2}{\widehat{\boldsymbol{\lambda}}'\mathbf{S}_p\widehat{\boldsymbol{\lambda}}} \end{aligned} \tag{2.14}$$

donde $d = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

El valor máximo que alcanza el cociente poblacional en 2.3 es

$$\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Esto es la distancia al cuadrado, $\boldsymbol{\Delta}^2$, entre dos poblaciones. El máximo valor que alcanza el cociente muestral en 2.14 se obtiene haciendo $\widehat{\boldsymbol{\lambda}} = \mathbf{S}_p (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. Por lo tanto

$$\begin{aligned} \max_{\widehat{\lambda}} \frac{\left(\widehat{\boldsymbol{\lambda}}'\mathbf{d}\right)^2}{\widehat{\boldsymbol{\lambda}}'\mathbf{S}_p\widehat{\boldsymbol{\lambda}}} &= \mathbf{d}'\mathbf{S}_p^{-1}\mathbf{d} \\ &= (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = \mathbf{D}^2 \end{aligned} \tag{2.15}$$

donde $\mathbf{D}^2$ es el cuadrado de la distancia muestral.

Para dos poblaciones, la separación relativa máxima que se puede alcanzar considerando combinaciones lineales de observaciones multivariadas es igual a la distancia $\mathbf{D}$. Esto es conveniente porque $\mathbf{D}^2$ puede ser utilizada, en ciertas situaciones, para probar si las medias poblacionales $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$ difieren significativamente. En consecuencia, una prueba para diferencias de vectores de medias puede ser vista como una prueba para la "significancia" de la separación que puede ser alcanzada.

Es importante hacer notar que una separación significativa no necesariamente implica una buena clasificación. La eficacia de un procedimiento de clasificación puede ser evaluada independientemente de cualquier prueba de separación. Por otro lado, si la separación no es significativa, la búsqueda de una regla de clasificación tal vez sea en vano.

Para ilustrar los conceptos anteriormente descritos, veamos el siguiente ejemplo.

En este problema se desea encontrar un procedimiento para separar dos poblaciones de iris[1]. Para ello, se tomaron muestras de cuatro variables en cada población:

[1]El iris es un género de planta bulbosa, con vistosas flores, cuyas especies están distribuidas por todas las regiones templadas del hemisferio norte.

$$
\begin{aligned}
X_1 &= \text{Longitud de sépalo} \\
X_2 &= \text{Ancho de sépalo} \\
X_3 &= \text{Longitud de pétalo} \\
X_4 &= \text{Ancho de pétalo}
\end{aligned}
$$

El primer grupo de tamaño $n_1 = 50$ fue seleccionado de una población de iris setosa y el segundo grupo de tamaño $n_2 = 50$ fue seleccionado de una población de iris versicolor. En las figuras 1 y 2 se muestran las gráficas en donde se comparan los dos grupos de iris de acuerdo a las medidas de sépalo y pétalo respectivamente.

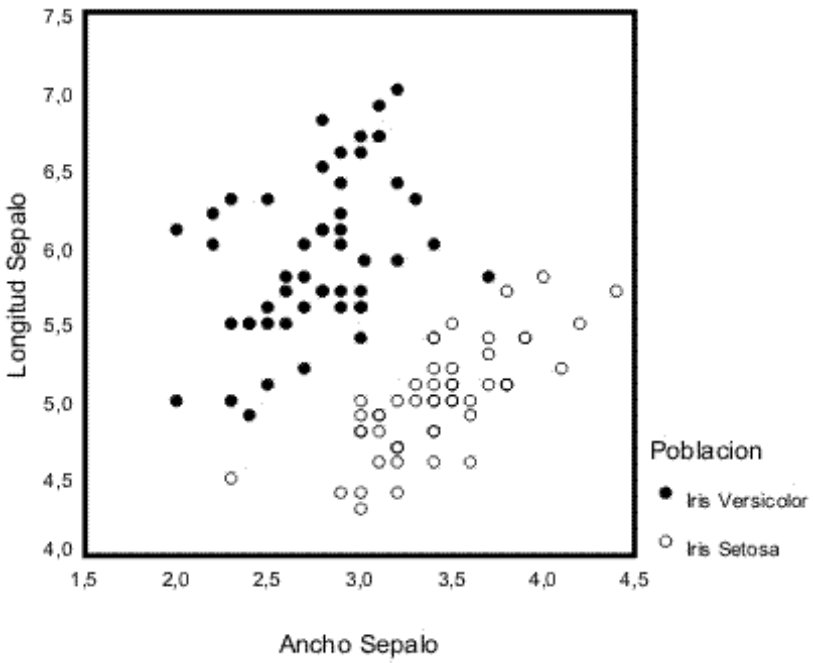

Tabla 2.2: Gráfica de las variables ancho y longitud de sépalo.

En la primera gráfica se puede observar que en genaral, la longitud de sépalo es mayor en la población de iris versicolor que en la de iris setosa,

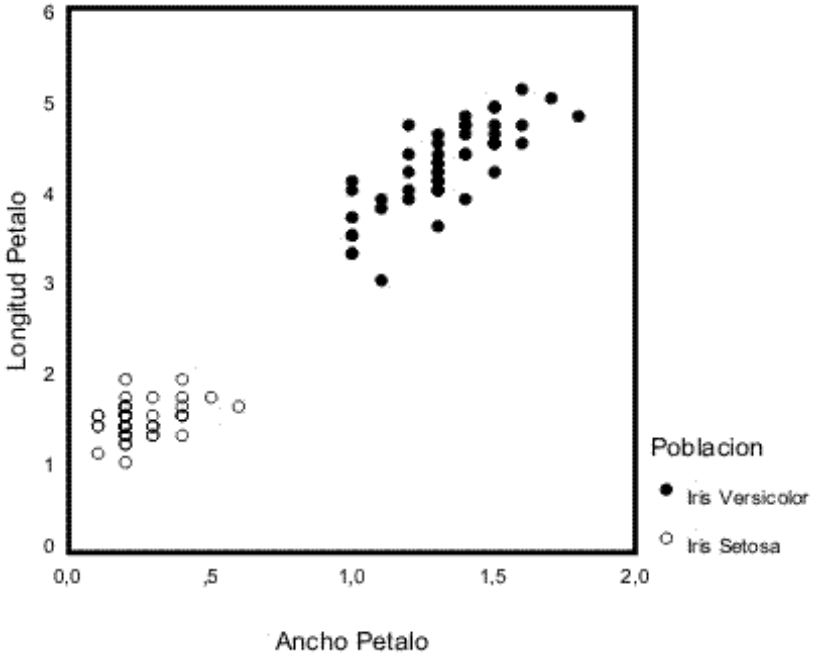

Tabla 2.3: Gráfica de las variables ancho y longitud de pétalo.

mientras que el ancho de sétalo es mayor en la población de iris setosa. En la segunda gráfica, la población de iris versicolor tiene mayor longitud y ancho de pétalo que la de iris setosa.

En ambas gráficas se puede notar que los dos grupos pueden ser "separados" sin mayor problema.

Al analizar los datos de las dos muestras, se obtiene la siguiente información:

$$\overline{\mathbf{X}}_1 = \begin{bmatrix} 5{,}006 \\ 3{,}428 \\ 1{,}462 \\ 0{,}246 \end{bmatrix} \qquad \overline{\mathbf{X}}_2 = \begin{bmatrix} 5{,}936 \\ 2{,}770 \\ 4{,}260 \\ 1{,}326 \end{bmatrix}$$

$$\mathbf{S}_p^{-1} = \begin{bmatrix} 11{,}09 & -5{,}22 & -7{,}59 & 1{,}746 \\ -5{,}22 & 11{,}06 & 2{,}37 & -6{,}15 \\ -7{,}59 & 2{,}37 & 21{,}22 & -25{,}49 \\ 1{,}74 & -6{,}15 & -25{,}49 & 83{,}14 \end{bmatrix}$$

La matriz de covarianzas común se tomó como el promedio ponderado de ambas matrices de covarianzas. De esta manera, la función discriminante de Fisher es

$$\begin{aligned} y &= \widehat{\boldsymbol{\lambda}}'\mathbf{X} = \left[\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right]' \mathbf{S}_p^{-1}\mathbf{X} \\ &= \begin{bmatrix} -0{,}93 \\ 0{,}65 \\ -2{,}79 \\ -1{,}08 \end{bmatrix}' \begin{bmatrix} 11{,}09 & -5{,}22 & -7{,}59 & 1{,}746 \\ -5{,}22 & 11{,}06 & 2{,}37 & -6{,}15 \\ -7{,}59 & 2{,}37 & 21{,}22 & -25{,}49 \\ 1{,}74 & -6{,}15 & -25{,}49 & 83{,}14 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \\ &= 5{,}617X_1 + 12{,}148X_2 - 23{,}218X_3 - 24{,}138X_4 \end{aligned}$$

Por otro lado

$$\overline{y}_1 = \widehat{\boldsymbol{\lambda}}'\mathbf{X}_1 = \begin{bmatrix} 5{,}617 & 12{,}148 & -23{,}218 & -24{,}138 \end{bmatrix} \begin{bmatrix} 5{,}006 \\ 3{,}428 \\ 1{,}462 \\ 0{,}246 \end{bmatrix} = 27{,}697$$

$$\overline{y}_2 = \widehat{\boldsymbol{\lambda}}'\mathbf{X}_2 = \begin{bmatrix} 5{,}617 & 12{,}148 & -23{,}218 & -24{,}138 \end{bmatrix} \begin{bmatrix} 5{,}936 \\ 2{,}770 \\ 4{,}260 \\ 1{,}326 \end{bmatrix} = -63{,}918$$

El punto medio entre $\overline{y}_1$ y $\overline{y}_2$ es

$$\widehat{m} = \frac{1}{2}\left(\overline{y}_1 + \overline{y}_2\right) = \frac{1}{2}\left(27{,}697 - 63{,}918\right) = -18{,}110$$

Ahora supongamos que tenemos una nueva observación $\mathbf{x}_0$ y queremos clasificarla en una de las dos poblaciones π_1: iris setosa o π_2: iris versicolor. En este caso, nuestra regla de asignación sería:

$$\begin{aligned} \text{Asigne } \mathbf{x}_0 \text{ a } \pi_1 \text{ si } y_0 &= \widehat{\boldsymbol{\lambda}}\mathbf{x}_0 \geq \widehat{m} = -18{,}110 \\ \text{Asigne } \mathbf{x}_0 \text{ a } \pi_2 \text{ si } y_0 &= \widehat{\boldsymbol{\lambda}}\mathbf{x}_0 < \widehat{m} = -18{,}110 \end{aligned}$$

2.3. El problema general de clasificación

En esta sección nos concentraremos en el problema de la clasificación para dos poblaciones, pero la mayoría de las ideas aquí presentadas pueden ser generalizadas para el caso de más de dos poblaciones.

Las reglas de asignación o clasificación usualmente son obtenidas a partir de muestras. Esencialmente, el conjunto de todas las posibles muestras es dividido en dos regiones, R_1 y R_2, tal que, si una nueva observación proviene de R_1, ésta será asignada a la población π_1, y si proviene de R_2, la asignaremos a π_2. Por lo tanto, un conjunto de valores observados favorece a π_1 y otro conjunto de valores favorece a π_2.

Un buen procedimiento de clasificación debería tener pocos errores cuando se ejecuta. En otras palabras, es deseable que las probabilidades de cometer un error de clasificación sean pequeñas. Como veremos más adelante, hay varias características que una regla de clasificación "óptima" debe tener.

Podría darse el caso en el que una población tuviera una probabilidad de ocurrencia mayor que otra debido a que una de las dos poblaciones es mucho más grande que la otra. Por ejemplo, una especie de planta A puede ser más abundante que otra especie B. Una regla de clasificación óptima debe tomar en cuenta este tipo de probabilidades de ocurrencia a priori. De esta manera, si uno piensa que la probabilidad (a priori) de ocurrencia de la planta B

es muy pequeña, una planta que es seleccionada aleatoriamente deberíamos clasificarla como tipo A, a menos que los datos favorezcan abrumadoramente al tipo B.

Otro aspecto de la clasificación es el costo. Supongamos que clasificar un objeto de la población π_1 en la población π_2 representa un error más grave que clasificar un objeto de la población π_2 en la población π_1. Entonces uno debe tener cuidado al hacer la anterior asignación. Por ejemplo, equivocarse al diagnosticar una enfermedad fatal es más costoso que decir que está presente cuando en realidad no está. Una clasificación óptima debe tomar en cuenta, siempre que sea posible, los costos asociados con los errores de clasificación.

Sean $f_1(x)$ y $f_2(x)$ las funciones de densidad de probabilidad asociadas con el vector aleatorio $\mathbf{X}$ para las poblaciones π_1 y π_2 respectivamente. Un objeto, con vector de medidas asociadas $\mathbf{x}$, debe ser asignado a π_1 o a π_2. Sea Ω el espacio muestral, esto es, la colección de todas las posibles observaciones $\mathbf{x}$. Sea R_1 el conjunto de valores de $\mathbf{x}$ para los cuales, clasificamos los objetos en π_1 y $R_2 = \Omega - R_1$ los valores restantes de $\mathbf{x}$ para los cuales clasificamos los objetos en π_2. Puesto que cada objeto debe ser asignado a una y solo una de las poblaciones, los conjuntos R_1 y R_2 son mutuamente excluyentes. Para el caso $p = 2$, podríamos tener un caso como el de la figura 2.4

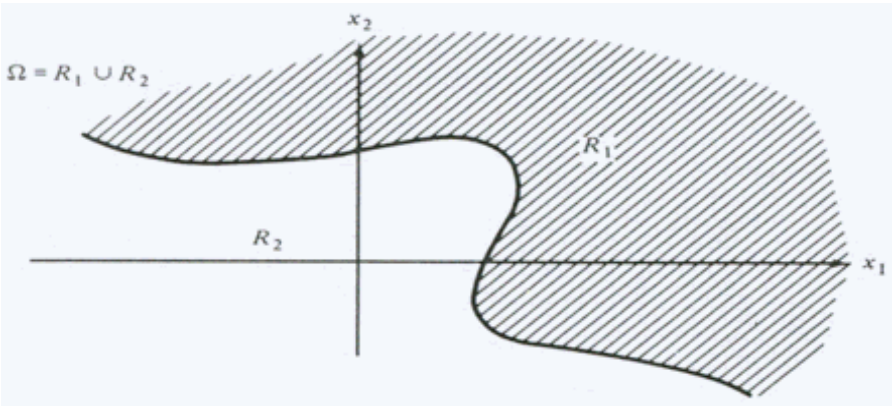

Tabla 2.4: Regiones de clasificación para dos poblaciones.

La probabilidad condicional, $P(2 \mid 1)$, de clasificar un objeto en π_2 cuando en realidad proviene de π_1 es

$$P(2 \mid 1) = P(\mathbf{X} \in \mathbf{R}_2 \mid \pi_1) = \int_{R_2=\Omega-R_1} f_1(\mathbf{x})\, d\mathbf{x} \tag{2.16}$$

Similarmente, la probabilidad condicional, $P(1 \mid 2)$, de clasificar un objeto en π_1 cuando en realidad proviene de π_2 es

$$P(1 \mid 2) = P(\mathbf{X} \in \mathbf{R}_1 \mid \pi_2) = \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \tag{2.17}$$

La integral en 2.16 representa el volumen formado por la función de densidad $f_1(\mathbf{x})$ sobre la región R_2. Similarmente, la integral en 2.17 representa el volumen formado por $f_2(\mathbf{x})$ sobre la región R_1. Esto se ilustra en la figura 2.5 para el caso univariado, $p = 1$.

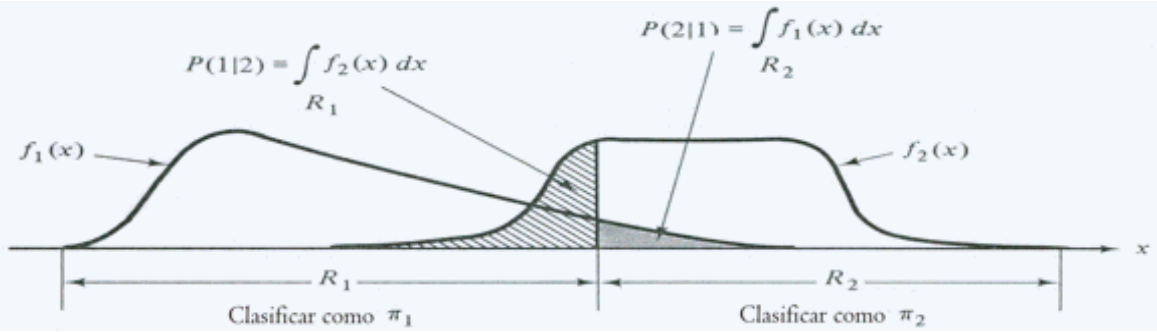

Tabla 2.5: Probabilidad de cometer error de clasificación para regiones de clasificación hipotéticas cuando p=1

Sea p_1 la probabilidad a priori de π_1 y p_2 la probabilidad a priori de π_2, donde $p_1 + p_2 = 1$. La probabilidad de clasificar un objeto correcta o incorrectamente puede ser derivada del producto de las probabilidades de clasificación a priori y condicionales:

$$\begin{aligned} & P(\text{objeto correctamente clasificado en } \pi_1) \\ = \; & P\left(\begin{array}{c}\text{observación proviene de } \pi_1 \\ \text{y es correctamente clasificada en } \pi_1\end{array}\right) \\ = \; & P(\mathbf{X} \in \mathbf{R}_1 \mid \pi_1)\, P(\pi_1) = P(1 \mid 1)\, p_1 \end{aligned}$$

$$
\begin{aligned}
& P\left(\text{objeto incorrectamente clasificado en } \pi_1\right) \\
= \; & P\left(\begin{array}{c}\text{observación proviene de } \pi_2 \\ \text{y es incorrectamente clasificada en } \pi_1\end{array}\right) \\
= \; & P\left(\mathbf{X} \in \mathbf{R}_1 \mid \pi_2\right) P\left(\pi_2\right) = P\left(1 \mid 2\right) p_2
\end{aligned}
$$

$$
\begin{aligned}
& P\left(\text{objeto correctamente clasificado en } \pi_2\right) \\
= \; & P\left(\begin{array}{c}\text{observación proviene de } \pi_2 \\ \text{y es correctamente clasificada en } \pi_2\end{array}\right) \\
= \; & P\left(\mathbf{X} \in \mathbf{R}_2 \mid \pi_2\right) P\left(\pi_2\right) = P\left(2 \mid 2\right) p_2
\end{aligned}
$$

$$
\begin{aligned}
& P\left(\text{objeto incorrectamente clasificado en } \pi_2\right) \\
= \; & P\left(\begin{array}{c}\text{observación proviene de } \pi_1 \\ \text{y es incorrectamente clasificada en } \pi_2\end{array}\right) \\
= \; & P\left(\mathbf{X} \in \mathbf{R}_2 \mid \pi_1\right) P\left(\pi_1\right) = P\left(2 \mid 1\right) p_1
\end{aligned}
\tag{2.18}
$$

Los esquemas de clasificación son a menudos evaluados en términos de sus probabilidades de error de clasificación, pero éstos ignoran los costos de clasificación. Por ejemplo, aún una probabilidad aparentemente pequeña como $0.06 = P\left(2 \mid 1\right)$, puede ser muy grande si el costo de hacer una asignación incorrecta en π_2 es extremadamente alto. Una regla que ignora los costos puede causarnos problemas.

Los costos de error de clasificación pueden ser definidos por una matriz de costos como la siguiente:

		Clasificada en	
		π_1	π_2
Población verdadera	π_1	0	$c\left(2 \mid 1\right)$
	π_2	$c\left(1 \mid 2\right)$	0

(2.19)

Los costos son: cero para una clasificación correcta, $c\,(1 \mid 2)$ cuando una observación que pertenece a π_2 es incorrectamente clasificada en π_1, y $c\,(2 \mid 1)$ cuando una observación de π_1 es incorrectamente clasificada en π_2.

Para cualquier regla, el costo promedio de cometer un error de clasificación (ECM, por sus siglas en inglés) se obtiene al multiplicar las entradas que están fuera de la diagonal de la matriz 2.19 por sus probabilidades de ocurrencia, obtenidas de 2.18 y en consecuencia

$$ECM = c\,(2 \mid 1)\,P\,(2 \mid 1)\,p_1 + c\,(1 \mid 2)\,P\,(1 \mid 2)\,p_2 \tag{2.20}$$

Una regla de clasificación razonable debería tener un ECM lo más pequeño posible. En la siguiente sección presentaremos reglas de este tipo.

2.4. Reglas de clasificación óptimas para dos poblaciones

Hemos sugerido que una regla de clasificación razonable podría ser determinada minimizando ECM. En otras palabras, las regiones de asignación R_1 y R_2 deben ser elegidas de tal manera que el ECM sea lo más pequeño posible.

Resultado 1. Las regiones R_1 y R_2 que minimizan el ECM están definidas por los valores de $\mathbf{x}$ para los cuales la siguiente desigualdad se cumple.

$$\begin{gathered}
R_1 : \frac{f_1\,(\mathbf{x})}{f_2\,(\mathbf{x})} \geq \left[\frac{c\,(1 \mid 2)}{c\,(2 \mid 1)}\right]\left[\frac{p_2}{p_1}\right] \\
\left[\begin{array}{c}\text{Cociente de}\\ \text{Densidades}\end{array}\right] \geq \left[\begin{array}{c}\text{Cociente de}\\ \text{Costos}\end{array}\right]\left[\begin{array}{c}\text{Cociente de}\\ \text{Probabilidades a Priori}\end{array}\right] \\
R_2 : \frac{f_1\,(\mathbf{x})}{f_2\,(\mathbf{x})} < \left[\frac{c\,(1 \mid 2)}{c\,(2 \mid 1)}\right]\left[\frac{p_2}{p_1}\right] \\
\left[\begin{array}{c}\text{Cociente de}\\ \text{Densidades}\end{array}\right] < \left[\begin{array}{c}\text{Cociente de}\\ \text{Costos}\end{array}\right]\left[\begin{array}{c}\text{Cociente de}\\ \text{Probabilidades a Priori}\end{array}\right]
\end{gathered} \tag{2.21}$$

Prueba. Sustituyendo las integrales para $P(2 \mid 1)$ y $P(1 \mid 2)$ dadas por 2.16 y 2.17 en 2.20 da como resultado

$$ECM = c(2 \mid 1)\, p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + c(1 \mid 2)\, p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

Notemos que $\Omega = R_1 \cup R_2$, entonces la probabilidad total

$$1 = \int_{\Omega} f_1(\mathbf{x})\, d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x}$$

de esta manera, podemos escribir

$$ECM = c(2 \mid 1)\, p_1 \left[1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x}\right] + c(1 \mid 2)\, p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

Luego, por la propiedad aditiva de las integrales,

$$ECM = \int_{R_1} \left[c(1 \mid 2)\, p_2 f_2(\mathbf{x}) - c(2 \mid 1)\, p_1 f_1(\mathbf{x})\right] d\mathbf{x} + c(2 \mid 1)\, p_1$$

Ahora, p_1, p_2, $c(1 \mid 2)$ y $c(2 \mid 1)$ son cantidades no negativas. Además, $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ son no negativas para toda $\mathbf{x}$ y son las únicas cantidades en el ECM que dependen de $\mathbf{x}$. Por lo tanto el ECM es minimizado si R_1 incluye aquellos valores $\mathbf{x}$ para los cuales el integrando

$$\left[c(1 \mid 2)\, p_2 f_2(\mathbf{x}) - c(2 \mid 1)\, p_1 f_1(\mathbf{x})\right] \leq 0$$

y excluye aquellas $\mathbf{x}$ para las cuales ésta cantidad es positiva. Esto es, R_1 debe ser el conjunto de puntos $\mathbf{x}$ tales que

$$c(1 \mid 2)\, p_2 f_2(\mathbf{x}) \leq c(2 \mid 1)\, p_1 f_1(\mathbf{x})$$

o

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left[\frac{c(1 \mid 2)}{c(2 \mid 1)}\right] \left[\frac{p_2}{p_1}\right]$$

Puesto que R_2 es el complemento de R_1 en Ω, R_2 debe ser el conjunto de punto para los cuales

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left[\frac{c(1 \mid 2)}{c(2 \mid 1)}\right] \left[\frac{p_2}{p_1}\right]$$

Es claro, de 2.21, que la regla del ECM mínimo requiere: (1) el cociente de las funciones de densidad evaluado en una nueva observación $\mathbf{x}$, (2) el cociente de costos, y (3) el cociente de probabilidades a priori. A menudo es mucho más fácil especificar los valores de los cocientes que los de sus componentes.

Por ejemplo, resultaría difícil especificar los costos (en unidades apropiadas) de clasificar a un alumno de una escuela de manejo como apto para manejar, cuando en realidad no lo es y clasificar a un alumno no apto para manejar cuando en realidad lo es. Sin embargo se puede dar el caso en el que se obtenga un número realista para el cociente de costos. Cualesquiera que sean las unidades de medida, clasificar a un alumno como apto para manejar, cuando en realidad no lo es, puede ser cinco veces más costoso que clasificar a un alumno como no apto, cuando en realidad si es. En este caso, el cociente de costos es cinco.

Resulta interesante considerar las regiones de clasificación definidas en 2.21 para algunos casos especiales.

1. $(p_2/p_1) = 1$(Probabilidades a priori iguales)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left[\frac{c(1 \mid 2)}{c(2 \mid 1)}\right]; \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left[\frac{c(1 \mid 2)}{c(2 \mid 1)}\right]$$

2. $[c(1 \mid 2)/c(2 \mid 1)] = 1$(costos de error de clasificación iguales)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_1}{p_2}; \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_1}{p_2}$$

3. $[p_2/p_1] = [c(1 \mid 2)/c(2 \mid 1)] = 1$(probabilidades a priori iguales y costos de error de clasificación iguales)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1; \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Cuando las probabilidades a priori son desconocidas, se toman como iguales y la regla del ECM mínimo se reduce a comparar el cociente de las

densidades de población con el cociente de costos de error de clasificación. Si el costo de cometer un error de clasificación es desconocido, usualmente se toma como la unidad y el cociente de densidades de población se compara con el cociente de probabilidades a priori. Finalmente, cuando los cocientes de probabilidades a priori y de costos de error de clasificación son iguales a la unidad, o en otras palabras, uno de los cocientes es el recíproco del otro, la región de clasificación óptima se determina simplemente comparando los valores de las funciones de densidad. En este caso, si $\mathbf{x}_0$ es una nueva observación y $f_1(\mathbf{x}_0)/f_2(\mathbf{x}_0) \geq 1$ [esto es $f_1(\mathbf{x}_0) \geq f_2(\mathbf{x}_0)$], asignamos $\mathbf{x}_0$ a π_1. Por otro lado, si $f_1(\mathbf{x}_0)/f_2(\mathbf{x}_0) < 1$ [es decir $f_1(\mathbf{x}_0) < f_2(\mathbf{x}_0)$], asignamos $\mathbf{x}_0$ a π_2.

Es práctica común usar arbitrariamente el caso (c) para clasificación. Esto es equivalente a suponer probabilidades a priori iguales y costos de error de clasificación iguales para la regla de ECM mínimo. Las reglas de asignación apropiadas para los casos que involucran probabilidades a priori iguales y costos de error de clasificación iguales, corresponden a funciones designadas para maximizar la separación de las poblaciones. Es en esta situación donde se empieza a notar la diferencia entre discriminación y clasificación.

Un criterio diferente al del costo esperado de error de clasificación puede ser usado para deducir un procedimiento de clasificación óptimo. Por ejemplo, podríamos ignorar los costos de error de clasificación y elegir R_1 y R_2 de tal manera que minimicen la *probabilidad total de error de clasificación* (TPM).

$TPM = P$(probabilidad de que la observación provenga de π_1, y sea clasificada en π_2)+P(probabilidad de que la observación provenga de π_2, y sea clasificada en π_1) es igual a

$$p_1 \int_{R_2} f_1(\mathbf{x})\, dx + p_2 \int_{R_1} f_2(\mathbf{x})\, dx \tag{2.22}$$

Matemáticamente, este problema es equivalente a minimizar el ECM, cuando los costos de error de clasificación son iguales. En consecuencia, las

regiones óptimas para este caso son las siguientes:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_1}{p_2}; \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_1}{p_2}$$

2.5. Clasificación con dos poblaciones normales multivariadas

Ahora supongamos que $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ son dos funciones de densidad normales multivariadas; la primera con vector de medias μ_1 y matriz de covarianzas Σ_1 y la segunda con vector de medias μ_2 y matriz de covarianzas Σ_2.

Sea $\Sigma_1 = \Sigma_2$. La función discriminante lineal de Fisher puede ser utilizada en este caso para clasificación ya que fue desarrollada bajo la hipótesis de que las dos poblaciones, cualquiera que sea su función de densidad, tienen la misma matriz de covarianzas. Por consiguiente, no es de sorprenderse que el método de Fisher corresponde a un caso particular de la regla del costo esperado de error de clasificación mínimo, la cual desarrollaremos a continuación.

Sea $f_i(\mathbf{x})$ $i = 1, 2$ las densidades conjuntas de $\mathbf{X}' = [X_1, X_2, ..., X_p]$ para las poblaciones π_1 y π_2, donde

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\sum|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right] \tag{2.23}$$

para $i = 1, 2$. Supongamos que los parámetros $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y Σ son conocidos.

Después de la cancelación de los términos $(2\pi)^{p/2}$ y $|\sum|^{1/2}$, y reordenando los exponentes de las densidades normales multivariadas en 2.23, la regiones en 2.21 que minimizan ECM son

$$R_1 :$$

$$\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)\right] \geq \left[\frac{c(1 \mid 2)}{c(2 \mid 1)}\right] \left[\frac{p_2}{p_1}\right]$$

$$R_2 :$$

$$\exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)'\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)\right] < \left[\frac{c\left(1 \mid 2\right)}{c\left(2 \mid 1\right)}\right]\left[\frac{p_2}{p_1}\right] \tag{2.24}$$

Dadas las regiones R_1 y R_2, podemos construir la siguiente regla de clasificación.

Resultado 2. Sean π_1 y π_2 poblaciones descritas por densidades normales multivariadas de la forma 2.23. La regla de asignación que minimiza el ECM es como sigue:

Asigne $\mathbf{x}_0$ a π_1 si:

$$\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\mathbf{x}_0-\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right) \geq \ln\left[\frac{c\left(1 \mid 2\right)}{c\left(2 \mid 1\right)}\right]\left[\frac{p_2}{p_1}\right] \tag{2.25}$$

Asigne $\mathbf{x}_0$ a π_2 en otro caso.

Prueba. Puesto que las cantidades en 2.24 son no negativas para todo $\mathbf{x}$, podemos tomar logaritmos naturales y las desigualdades se mantienen. Más aún

$$\begin{aligned} &-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)'\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_2\right) \\ &\quad=\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right) \end{aligned} \tag{2.26}$$

y en consecuencia

$$\begin{aligned} & & R_1 \quad : & & \\ \left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right) & \geq & \ln\left[\frac{c\left(1 \mid 2\right)}{c\left(2 \mid 1\right)}\right]\left[\frac{p_2}{p_1}\right] & \\ & & R_2 \quad : & & \\ \left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\Sigma^{-1}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right) & < & \ln\left[\frac{c\left(1 \mid 2\right)}{c\left(2 \mid 1\right)}\right]\left[\frac{p_2}{p_1}\right] & \end{aligned} \tag{2.27}$$

En la mayoría de las situaciones prácticas, las cantidades poblacionales $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y Σ son desconocidas, por lo que la regla 2.25 debe ser modificada. Algunos autores han sugerido sustituir los parámetros poblacionales por sus

contrapartes muestrales. Al sustituir $\overline{\mathbf{x}}_1$ por $\boldsymbol{\mu}_1$, $\overline{\mathbf{x}}_2$ por $\boldsymbol{\mu}_2$ y $\mathbf{S}_p$ por Σ en 2.25, se obtiene la siguiente regla de clasificación "muestral":

Asigne $\mathbf{x}_0$ a π_1 si

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' S_p^{-1} \mathbf{x}_0 - \frac{1}{2} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' S_p^{-1} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \geq \ln \left[\left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_1}{p_2} \right) \right] \tag{2.28}$$

asigne $\mathbf{x}_0$ a π_2 en otro caso.

La expresión

$$\begin{aligned} w &= (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' S_p^{-1} \mathbf{x} - \frac{1}{2} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' S_p^{-1} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \\ &= (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' S_p^{-1} \left[\mathbf{x} - \frac{1}{2} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \right] \end{aligned} \tag{2.29}$$

es a menudo llamada *la función de clasificación de Anderson.*
Si $[(c(1 \mid 2)/c(2 \mid 1))(p_2/p_1)] = 1$, es decir, $\ln[(c(1 \mid 2)/c(2 \mid 1))(p_2/p_1)] = 0$, la regla 2.28 es comparable con la regla 2.13 basada en la función discriminante lineal de Fisher. Por lo tanto, si las dos poblaciones tienen la misma matriz de covarianzas, la regla de clasificación de Fisher es equivalente a la regla del mínimo ECM cuando los costos de error de clasificación y las probabilidades a priori son iguales.

Una vez que los parámetros estimados son introducidos en lugar de las correspondientes cantidades poblacionales, no hay seguridad de que la regla resultante minimizará el costo de error de clasificación esperado en un aplicación particular, ya que la regla óptima en 2.25 fue obtenida suponiendo que las densidades normales multivariadas $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ estaban completamente determinadas. La expresión 2.28 es simplemente un estimador de la regla óptima. Sin embargo, es razonable pensar que debería funcionar bien si el tamaño de muestra es grande.

En resumen, si los datos tienen una distribución normal multivariada, la estadística w en 2.29 puede ser calculada para cada nueva observación $\mathbf{x}_0$. Estas observaciones son clasificadas comparando los valores de w con el valor de $\ln[(c(1 \mid 2)/c(2 \mid 1))(p_2/p_1)]$ como en 2.28.

Capítulo 3

La aplicación

3.1. Introducción

Como ya lo vimos en los capítulos 1 y 2, con regresión logística y análisis discriminante podemos dar respuesta al problema de clasificar un objeto o individuo como miembro de una de dos poblaciones con base en un conjunto de medidas previamente registradas. Una vez revisada la teoría del análisis de regresión logística y análisis discriminante, vamos a aplicar ambas técnicas a un problema en particular y compararemos los resultados en términos del porcentaje de clasificaciones correctas que hace cada uno de los métodos.

Los datos para este ejercicio fueron tomados de un estudio realizado por el Baystate Medical Center de Springfield, Massachusetts en el año 1986. El conjunto de datos completo puede encontrarse en el apéndice A.

Se trata de un estudio que pretende identificar los factores de riesgo relacionados a los bebés que nacen con bajo peso (2500 gr. o menos). Los nacimientos de bebés con bajo peso han sido objeto de estudio de médicos e investigadores por muchos años. Esto se debe a que la tasa de mortalidad infantil y la tasa de defectos congénitos son muy altas para los bebés nacidos con bajo peso. El comportamiento de la mujer durante el embarazo (dieta, hábito de fumar, atención médica prenatal, etc.) pueden alterar de manera importante las posibilidades de llevar el embarazo hasta su término

y consecuentemente dar a luz a un bebé con peso normal.

Para este análisis se recolectaron datos de 189 mujeres, 59 de las cuales tuvieron bebés de bajo peso y 130 tuvieron bebés con peso normal. Los datos consisten en las mediciones de 9 variables presumiblemente relacionadas con la condición de bajo peso: edad de la madre en años (AGE), peso en libras en el último periodo menstrual (LWT), raza (RACE), status fumador durante el embarazo (SMOKE), número de partos prematuros (PTL), registro de hipertensión (HT), presencia de irritabilidad uterina (IU), número de visitas al médico durante el primer trimestre del embarazo (FTV) y peso de la madre al nacer, en gramos (BWT). La variable dependiente será el peso al nacer del bebé (LOW) que es la condición que queremos relacionar a las variables anteriormente descritas.

Notemos que algunas de las variables independientes son variables categóricas, es decir, los valores que pueden tomar sólo indican la pertenencia a cierto grupo. Estas variables y sus respectivas codificaciones son: raza (1 = blanca, 2 = negra, 3 = otra), status fumador durante el embarazo (1 = si, 0 = no), número de partos prematuros (0 = ninguno, 1 = uno, etc.), registro de hipertensión (0 = no, 1 = si), presencia de irritabilidad uterina (0 = no, 1 = si), y número de visitas durante el primer trimestre de embarazo (0 = ninguno, 1 = uno, etc.).

Una vez planteado el problema ejecutamos el análisis estadístico en SPSS v10.0 y describiremos los resultados en cada caso. En principio analizaremos los datos mediante regresión logística.

3.2. Regresión logística

Como primer paso, se hizo un análisis previo para seleccionar las variables que tuvieran mayor poder de discriminación. De esta manera, con base en el método de selección por pasos[1] (stepwise), las más significativas resultaron

[1]El método de selección por pasos o *stepwise* sirve para definir las variables que deben ser incluídas en un modelo, con base en criterios estadísticos.

ser: peso durante el último periodo menstrual, raza, presencia de irritabilidad uterina, historia de hipertensión y status fumador durante el embarazo. A partir de este momento, el desarrollo del modelo lo haremos usando sólo estas variables.

Resumen de los casos procesados

Casos no ponderados		N	Percent
Casos seleccionados	Incluídos	189	100.0
	Excluídos	0	.0
	Total	189	100.0
Casos no seleccionados		0	.0
Total		189	100.0

Tabla 3.1:

La tabla 3.1 muestra un resumen del procesamiento de los casos. En ella podemos ver que el 100 % de los 189 casos se procesaron correctamente. Dentro de los casos excluidos se contemplan aquellos cuya variable de agrupación no está dentro del rango seleccionado o tienen un valor perdido en al menos una variable discriminante. En nuestro ejemplo no hay casos excluidos.

Codificación de la variable dependiente

Valor Original	Valor Interno
No	0
Si	1

Tabla 3.2:

La tabla 3.2 nos indica que la variable dependiente, LOW, se codifica

como 0 si el recién nacido no presenta bajo peso y 1 si presenta bajo peso. En este caso el análisis se enfocará en la probabilidad de que la variable LOW sea igual a 1, es decir, la probabilidad de que el recién nacido presente bajo peso.

Codificación de las variables categóricas

			Código del parámetro	
		Frecuencia	(1)	(2)
Raza	Blanca	96	1.000	.000
	Negra	26	.000	1.000
	Otra	67	.000	.000
Presencia de irritabilidad uterina	No	161	1.000	
	Si	28	.000	
Historia de Hipertensión	No	177	1.000	
	Si	12	.000	
Fumó durante el embarazo	No	115	1.000	
	Si	74	.000	

Tabla 3.3:

SPSS genera las variables dummy que se necesitarán para el manejo de las variables categóricas. Notemos en la tabla 3.3, que los parámetros para los coeficientes de la última categoría en cada variable son cero, lo cual nos indica que esta categoría será el valor omitido para el conjunto de variables dummy. Supongamos que llamamos D_1 y D_2 a las variables dummy generadas para la variable raza, entonces si nos referimos a la raza blanca $D_1 = 1$ y $D_2 = 0$, si aludimos a la raza negra $D_1 = 0$ y $D_2 = 1$ y si hablamos de otra raza entonces $D_1 = 0$ y $D_2 = 0$. En el caso de las variables presencia de irritabilidad uterina, historia de hipertensión y status fumador, sólo se generará una variable dummy que tendrá valor 1 si la característica no está

presente y 0 si está.

3.2.1. El modelo nulo

Como un primer paso, generamos el modelo que incluye sólo una constante y ninguna variable predictiva.

Tabla de Clasificación [a,b]

			Pronosticados		
			Bajo peso al nacer		Porcentaje
	Observados		No	Si	correcto
Paso 0	Bajo peso al nacer	No	130	0	100.0
		Si	59	0	.0
	Porcentaje total				68.8

a. Constante incluída en el modelo

b. El valor de corte es .500

Tabla 3.4:

La tabla de 3.4 clasifica como correctos o incorrectos los resultados estimados por el modelo nulo, es decir, aquel que sólo incluye la constante y ninguna variable dependiente. En las columnas se muestran los dos valores pronosticados de la variable dependiente, mientras que en los renglones están los valores reales. Bajo un modelo perfecto, todos los casos deberían estar incluidos en la diagonal de la tabla y porcentaje de clasificación correcta debería ser del 100 %. En este caso el modelo pronostica correctamente el 100 % de los casos que no presentan bajo peso al nacer contra 0 % de los casos que presentaron bajo peso al nacer. En promedio clasifica correctamente el 68.8 % de los casos totales.

Variables en la ecuación

		B	S.E.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-.790	.157	25.327	1	.000	.454

Tabla 3.5:

La tabla 3.5 muestra los resultados de la prueba de significancia del modelo en el cual los coeficientes de todas las variables independientes son cero. Tener un nivel de significancia de .000 como en nuestro caso, implica que el modelo nulo debe ser rechazado, y por lo tanto descartado.

3.2.2. El modelo completo

Variables en la ecuación

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	gl	Sig.	Exp(B)	Inferior	Superior
Paso 1[a]	LWT	-.017	.007	6.049	1	.014	.983	.970	.997
	RACE			8.117	2	.017			
	RACE(1)	-.926	.430	4.631	1	.031	.396	.170	.921
	RACE(2)	.398	.537	.551	1	.458	1.489	.520	4.264
	SMOKE(1)	-1.036	.393	6.962	1	.008	.355	.164	.766
	HT(1)	-1.871	.691	7.337	1	.007	.154	.040	.596
	UI(1)	-.905	.448	4.089	1	.043	.405	.168	.973
	Constante	4.795	1.30	13.548	1	.000	120.867		

a. Variables ingresadas en el paso 1: LWT, RACE, SMOKE, HT, UI.

Tabla 3.6:

Una vez descartada la posibilidad de ajuste con el modelo nulo, evaluaremos el problema con el modelo completo, es decir, integrando todas aquellas

variables presumiblemente significativas. La tabla 3.6 muestra los coeficientes de las variables independientes, las estadísticas de Wald y sus respectivos niveles de significancia para cada una de las variables, incluyendo las variables *dummy*. De esta manera tenemos que el modelo ajustado queda como sigue:

$$\begin{aligned}\widehat{g} &= 4.795 - 0.017LWT - 0.926RACE\_1 + 0.398RACE\_2 - \\ &\quad 1.036SMOKE\_1 - 1.871HT\_1 - 0.905UI\_1\end{aligned}$$

donde $\widehat{g}$ es la probabilidad estimada de que un bebé nazca con bajo peso.

Bajo la hipótesis de que los coeficientes son cero, la estadística de Wald tiene una distribución normal estándar como ya lo vimos en el capítulo 2. Los p-values se muestran en la sexta columna del cuadro anterior. Si usamos un nivel de significancia de 0.05 entonces concluiríamos que todas las variables son significativas excepto por la variable RACE_2, sin embargo, siempre que una variable categórica es incluída (o excluída) de un modelo, todas sus variables dummy deben ser incluidas (o excluidas) pues hacer lo contrario implicaría que estamos haciendo una recodificación de la variable.

La columna Exp(B) representa el cociente de momios (odds ratio[2]) de cada una de las variables independientes respecto a la variable dependiente. Exp(B) representa el cambio pronosticado en el cociente de momios por cada unidad en que se incremente la variable independiente correspondiente. Si el valor es mayor que uno, el cociente de momios se incrementa, si el valor es menor que uno, decrece. Un valor igual a uno no mueve el cociente de momios. En nuestro caso, El valor Exp(B) para la variable LWT es de 0.98, lo cual quiere decir que por cada libra adicional de peso antes del embarazo, las posibilidades de tener un bebé de bajo peso disminuyen a razón de 0.98, es decir, una mujer que pesa 120 libras es 9.8 veces menos probable que

[2]El *odds ratio*, también llamado cociente de momios, es el cociente entre la probabilidad de que un evento suceda y la probabilidad de que no suceda.

tenga un bebé con bajo peso al nacer que una mujer que pesa 110 libras. La afirmación anterior se cumple en el supuesto de que los demás parámetros se mantienen fijos.

Una pregunta interesante que surge en este punto es, ¿cuán mayor es la posibilidad de que una mujer de raza negra de a luz un bebé con bajo peso, comparada con una mujer de raza blanca? Para responder esta pregunta observemos que para el caso de una mujer de raza negra (RACE_2), Exp(B) = 1.489 y para el caso de una mujer de raza blanca (RACE_1), Exp(B) = 0.396. El cociente de estas dos cantidades 1.489/0.396 = 3.76 se interpreta como el número de veces que es más probable que una mujer de raza negra tenga un bebé con bajo peso al nacer comparada con una mujer de raza blanca.

Prueba Omnibus de los coeficientes del modelo

		Chi-square	df	Sig.
Paso 1	Paso	30.455	6	.000
	Bloque	30.455	6	.000
	Modelo	30.455	6	.000

Tabla 3.7:

Al construir el modelo que incluye todas las variables en cuestión, se realiza la prueba de Omnibus para saber si el paso del modelo nulo al modelo con las variables tiene justificación. Una significancia de 0.05 o menos en la estadística "Paso" es suficiente para concluir que la inclusión de las variables está justificada. En este caso, como lo muestra la tabla 3.7, se justifica la inclusión de las variables en cuestión.

Para evaluar la bondad de ajuste de nuestro modelo, usaremos la prueba de Hosmer y Lemeshow[3]. La idea de esta prueba es calcular, para cada obser-

[3]Para mayor referencia de esta prueba, consultar Hosmer and Lemeshow (2000)

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrada	gl	Sig.
1	11.710	8	.165

Tabla 3.8:

vación del conjunto de datos, las probabilidades de la variable independiente que predice el modelo, agruparlas y calcular, a partir de ellas, las frecuencias esperadas y compararlas con las esperadas mediante la prueba χ^2. Por ejemplo, la frecuencia observada en el grupo que no tuvo bajo peso al nacer, para el quinto decil de riesgo, es 15. Este valor es obtenido de la suma de los valores observados para los 18 bebés pertenecientes a este grupo. De manera similar, la frecuencia estimada para este decil es 13.130, la cual es resultado de la suma de las probabilidades estimadas para estos 18 bebés. La frecuencia observada para el grupo de bajo peso al nacer es 18-15=3, y la frecuencia estimada es 18-13.130=4.870.

De esta manera, si la significancia de la prueba de Hosmer y Lemeshow es igual o menor que 0.05, rechazamos la hipótesis nula de que no hay diferencias significativas entre los datos observados y los datos pronosticados de la variable dependiente. Si la significancia es mayor que 0.05, no rechazamos la hipótesis nula de que no hay diferencias entre los datos. En nuestro caso, la tabla 3.8 nos muestra la significancia de 0.165, lo cual implica que el modelo se ajusta aceptablemente a nuestros datos. Esto no quiere decir que sea el mejor modelo, sólo que existe un ajuste significativo con las variables consideradas.

La tabla 3.9 muestra la cantidad de datos observados en cada decil, separados en dos grupos. Junto a cada uno de los datos observados se muestran los datos esperados de acuerdo al modelo. En general se puede decir que las diferencias no son significativas y que el ajuste del modelo es aceptable.

Finalmente, la tabla 3.10 nos muestra los datos clasificados como correctos

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Bajo peso al nacer = No		Bajo peso al nacer = Si		
		Observados	Pronosticados	Observados	Pronosticados	Total
Paso 1	1	18	17.747	1	1.253	19
	2	18	16.845	1	2.155	19
	3	12	15.701	7	3.299	19
	4	16	14.505	3	4.495	19
	5	15	13.130	3	4.870	18
	6	13	13.261	6	5.739	19
	7	10	13.321	10	6.679	20
	8	14	11.251	5	7.749	19
	9	9	8.575	10	10.425	19
		5	5.665	13	12.335	18

Tabla 3.9:

e incorrectos para el modelo completo, incluyendo la constante. Las columnas contienen las datos pronosticados mientras que los renglones contienen los datos observados. Si el modelo tuviera un ajuste perfecto, todos los datos deberían estar en la diagonal y el porcentaje de clasificación correcta sería del 100 %. En nuestro caso el modelo acierta al clasificar 138 de los datos, 117 en la categoría de peso normal y 21 en la categoría de peso bajo, y falla en 51 ocasiones, 13 pertenecientes a la categoría de peso normal y 38 en la categoría de bajo peso al nacer. El porcentaje total de clasificación correcta es del 73 %.

En resumen, el modelo propuesto clasifica correctamente al 73 % de los datos analizados, por lo tanto puede considerarse como un modelo significativamente bueno.

Una vez revisados los resultados que arroja el análisis de regresión logística, vamos a analizar estos mismos datos utilizando análisis discriminante y posteriormente compararemos los resultados entre ambas técnicas.

Tabla de clasificación [a]

			Pronosticados		
			Bajo peso al nacer		Porcentaje correcto
	Observados		No	Si	
Paso 1	Bajo peso al nacer	No	117	13	90.0
		Si	38	21	35.6
	Porcentaje total				73.0

a. El valor de corte es .500

Tabla 3.10:

3.3. Análisis discriminante

Tabla resumen de los casos procesados

Casos no ponderados		N	Porcentaje
Válidos		189	100.0
Excluídos	Por pertenecer a un grupo fuera de rango	0	.0
	Por tener valor perdido en al menos una variable discriminante	0	.0
	Por pertenecer a un grupo fuera de rango o por tener valorperdido en al menos una variable discriminante	0	.0
	Total	0	.0
Total		189	100.0

Tabla 3.11:

La tabla 3.11 nos muestra que el procesamiento del 100 % de los 189 casos se realizó correctamente. No se encontraron observaciones con valores perdidos o fuera de rango.

En la tabla 3.12 se muestran algunas estadísticas descriptivas como son

Estadísticas por grupo

Bajo peso al nacer		Media	Desv. estandar	N válido (según lista)	
				No ponderados	Ponderados
No	LWT	133.30	31.72	130	130.000
	RACE	1.76	.91	130	130.000
	SOMKE	.34	.48	130	130.000
	HT	3.85E-02	.19	130	130.000
	IU	.11	.31	130	130.000
Si	LWT	122.14	26.56	59	59.000
	RACE	2.03	.91	59	59.000
	SOMKE	.51	.50	59	59.000
	HT	.12	.33	59	59.000
	IU	.24	.43	59	59.000
Total	LWT	129.81	30.58	189	189.000
	RACE	1.85	.92	189	189.000
	SOMKE	.39	.49	189	189.000
	HT	6.35E-02	.24	189	189.000
	IU	.15	.36	189	189.000

Tabla 3.12:

la media y desviación estándar de cada uno de los grupos y del total de la población estudiada. Estas estadísticas nos darán un panorama general de los datos. Podemos ver por ejemplo, que el peso promedio de las mujeres que dieron a luz bebés con bajo peso es 11 libras menor que el peso promedio de aquellas mujeres que dieron a luz bebés con peso normal.

Es importante notar que en el análisis discriminante las variables categóricas son tratadas como variables continuas, lo cual puede conducirnos a resultados poco precisos. Esta es una de las desventajas al usar análisis discriminante.

Como se vio en el capítulo 2, una de las suposiciones que se hace en el

análisis discriminate es la igualdad de las matrices de varianzas y covarianzas de ambas poblaciones. La prueba de Box nos sirve para verificar si se cumple esta igualdad. En este caso la el p-value de 0.000 mostrado en la tabla 3.13, nos indica que la prueba es significativa por lo que podemos concluir que las matrices de varianzas y covarianzas de ambas poblaciones difieren, violando de esta manera la suposición del análisis discriminante.

Tabla de Resultados de la prueba M de Box

Box's M		44.924
F	Aprox.	2.890
	gl1	15
	gl2	54069.345
	Sig.	.000

Tabla 3.13:

También vimos que cuando el supuesto de igualdad entre ambas matrices no se cumple, se pueden combinar ambas para derivar en un sólo estimador insesgado de $\sum$, el cual aparece en la tabla 3.14

Matriz de varianzas-covarianzas combinada

		LWT	RACE	SMOKE	HT	IU
Covariance	LWT	913.049	-4.000	-.253	1.971	-1.359
	RACE	-4.000	.832	-.163	-2.405E-04	9.967E-03
	SMOKE	-.253	-.163	.235	-1.346E-03	6.112E-03
	HT	1.971	-2.4E-04	-1.346E-03	5.870E-02	-1.176E-02
	IU	-1.359	9.967E-03	6.112E-03	-1.176E-02	.124

Tabla 3.14: La matriz de covarianzas tiene 187 grados de libertad

La tabla de variables introducidas/eliminadas muestra un resumen de todos los pasos llevado a cabo en la construcción de la función discriminante y recuerda los criterios utilizados en la selección de las variables. En cada

Variables introducidas/eliminadas[a,b,c,d]

		Lambda de Wilk's							
						F exacta			
Paso	Introducidas	Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	LWT	.971	1	1	187	5.540	1	187	.020
2	HT	.932	2	1	187	6.786	2	186	.001
3	IU	.906	3	1	187	6.374	3	185	.000
4	SMOKE	.887	4	1	187	5.879	4	184	.000
5	RACE	.863	5	1	187	5.813	5	183	.000

En cada paso se introduce la variable que minimiza la lambda de Wilks global

a. El número máximo de pasos es 10

b. La F parcial mínima para entrar es 3.84

c. La F parcial máxima para salir en 2.71

d. El nivel de F, la tolerancia o el VIN son suficientes para continuar los cálculos

Tabla 3.15:

paso se informa de la variable que ha sido incorporada al modelo, y en su caso, de la variable o variables que han sido eliminadas. En nuestro caso, todos los pasos llevados a cabo han sido de incorporación de variables: en el primer paso, LWT; en el segundo caso, HT ; etc. Así hasta un total de 5 variables. En ninguno de los 5 pasos ha habido eliminación de variables. Si alguna de las variables previamente incorporadas hubiera sido eliminada en algún paso posterior, la tabla mostraría una columna adicional indicando tal circunstancia.

Las notas al pie de la tabla recuerdan las opciones establecidas para el análisis: La selección de variables se ha llevado a cabo utilizando el estadístico lambda de Wilks global, el número máximo de pasos permitidos es 10, el valor del estadístico F para incorporar variables es de 3.84 (criterio de entrada), el valor del estadístico F para eliminar variables es de 2.71 (criterio de salida)

y, por último, la nota *d* indica que se ha alcanzado alguno de los criterios de parada (los niveles del estadístico F, el criterio de tolerancia y la V mínima de Rao).

Puede observarse que el valor del estadístico lambda de Wilks va disminuyendo en cada paso, lo cual es síntoma de que, conforme se van incorporando variables en el modelo, los grupos van estando cada vez menos solapados. En la columna *F Exacta* se encuentra el valor transformado de la lambda de Wilks y su significación, la cual en todos los casos en menor que 0.05 por lo que las variables en cuestión pueden ser consideradas en el modelo.

Variables incluídas en el análisis

Paso		Tolerancia	F para eliminar	Lambda de Wilks
1	LWT	1.000	5.540	
2	LWT	.928	8.939	.977
	HT	.928	7.830	.971
3	LWT	.919	7.269	.942
	HT	.917	8.968	.950
	IU	.972	5.241	.932
4	LWT	.919	6.938	.920
	HT	.917	8.760	.929
	IU	.971	4.783	.910
	SMOKE	.999	4.076	.906
5	LWT	.894	4.724	.885
	HT	.915	7.866	.900
	IU	.970	4.310	.883
	SMOKE	.857	7.326	.897
	RACE	.838	5.032	.887

Tabla 3.16:

La tabla 3.16 se encuentra dividida por cada uno de los pasos. En cada paso se mencionan las variables incorporadas al modelo hasta ese momento y, para cada variable, el nivel de tolerancia, el valor del estadístico F que

permite valorar si la variable debe o no ser eliminada (F para eliminar) y la lambda de Wilks global que obtendríamos si se eliminara la variable del modelo.

Esta tabla permite valorar (mediante F y lambda) el efecto de la exclusión de cada variable y, (mediante el nivel de tolerancia) el grado de colinealidad existente entre las variables independientes. El nivel de tolerancia es la proporción de varianza de una variable independiente que no está explicada por el resto de las variables independientes. Puesto que las variables independientes de nuestro ejemplo no están muy relacionadas entre sí, la tolerancia disminuye ligeramente en el momento en que se incorpora una nueva variable en el modelo. En el paso 0 todas las variables tienen una tolerancia igual a 1, pues todavía no existen variables en el modelo. En el paso 1 la primera variable permanece en ese valor, pues al estar sola, no existen variables que puedan explicar nada de ella (ver tolerancia de la variable LWT en el paso 1). En el segundo paso, al incorporarse la variable HT, la tolerancia baja a 0.928, lo cual indica que existe poca correlación entre ambas variables.

Lambda de Wilks

Paso	Número de variables	Lambda	gl1	gl2	gl3	F exacta			
						Estadístico	gl1	gl2	Sig.
1	1	.971	1	1	187	5.540	1	187	1.96E-02
2	2	.932	2	1	187	6.786	2	186	1.43E-03
3	3	.906	3	1	187	6.374	3	185	3.91E-04
4	4	.887	4	1	187	5.879	4	184	1.79E-04
5	5	.863	5	1	187	5.813	5	183	5.23E-05

Tabla 3.17:

La tabla 3.17 muestra el estadístico lambda de Wilks global para el mode-

lo generado en cada paso, independientemente de que se haya optado por otro estadístico como método de selección de variables. Este estadístico permite valorar el grado de diferenciación entre los grupos tomando como referencia las variables independientes incluidas en cada paso. Toma valores entre 0 y 1 de forma que, cuanto más cerca de 0 esté, mayor es el poder discriminante de las variables consideradas y cuanto más cerca de 1, menor es dicho poder.

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	.159	100.0	100.0	.370

Tabla 3.18:

La tabla 3.18 contiene los autovalores y algunas estadísticas descriptivas multivariantes. El autovalor es el cociente entre la variación debida a las diferencias entre los grupos (medida mediante la suma de cuadrados intra-grupos) y la variación que se da dentro de cada grupo combinada en una única cantidad (medida mediante la suma de cuadrados intra grupos). Este estadístico se diferencia de la F del análisis de varianza multivariada en que no intervienen los grados de libertad. Su interés principal radica en que permite comparar cómo se distribuye la dispersión intra-grupos cuando existe más de una función. Aunque un autovalor tiene un mínimo de cero, no tiene un máximo, lo cual lo hace difícilmente interpretable por sí solo. Por esta razón se prefiere utilizar el estadístico lambda de Wilks, que se encuentra estrechamente relacionado con los autovalores.

La correlación canónica es la correlación entre la combinación lineal de las variables independientes (la función discriminante) y una combinación lineal de variables *indicador* (unos y ceros) que recogen la pertenencia de los sujetos a los grupos. En el caso de dos grupos, la correlación canónica es la correlación simple entre las puntuaciones discriminantes y una variable

con códigos 1 y 0 según cada caso pertenezca a un grupo o a otro. Una correlación canónica alta indica que las variables discriminantes permiten diferenciar entre los grupos.

El autovalor obtenido en nuestro caso está bastante próximo a 0 y la correlación canónica es moderada, por lo que podemos suponer que las variables discriminantes utilizadas no permiten distinguir muy bien entre los grupos.

Lambda de Wilks

Contraste de Funciones	Lambda de Wilks	Chi-cuadrada	gl	Sig.
1	.863	27.196	5	.000

Tabla 3.19:

La tabla 3.19 muestra el valor de la lambda de Wilks para el modelo final. Su significación se evalúa mediante una transformación chi-cuadrada.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función
	1
LWT	-.453
RACE	.483
SMOKE	.573
HT	.573
IU	.416

Tabla 3.20:

La tabla 3.20 nos muestra la matriz de coeficientes estandarizados. Estos coeficientes permiten valorar la contribución *neta* de cada variable a la función discriminante. Así, las variables que más contribuyen a diferenciar los

grupos son SMOKE y HT. De esta manera, si la madre fuma o tiene historia de hipertensión, la puntuación en la función discriminante será mayor y en consecuencia tendrá mayor tendencia a tener un bebé de bajo peso. La variable LWT, sin embargo, presenta un coeficiente negativo. Esto quiere decir que para las mujeres con puntuaciones iguales en las variables restantes, las que tienen un mayor peso antes de su último periodo menstrual, tendrán una menor puntuación en la función discriminante y en consecuencia sus probabilidades de tener un bebé de bajo peso serán menores.

Matriz de estructura

	Función
	1
LWT	-.432
IU	.430
SMOKE	.410
HT	.387
RACE	.349

Tabla 3.21:

En la matriz de estructura anterior se encuentran los coeficientes de co-rrelación *brutos* entre cada variable y la función discriminante. Como podemos ver, la variable que más ayuda a discriminar entre los grupos es LWT mientras que la que menos aporta a la discriminación es RACE.

La tabla 3.22 muestra los centroides de cada grupo los cuales indican que las mujeres que dan a luz bebés con bajo peso, tienen en general mayores puntuaciones que aquellas mujeres que dan a luz bebés con peso normal.

La tabla de probabilidades previas o probabilidades *a priori* contiene las probabilidades asignadas a cada grupo. Estas probabilidades reflejan el tamaño relativo de cada grupo y en caso de que una nueva observación se encuentre equidistante a los centroides de los grupos, ésta se asignará al grupo que tenga la mayor probabilidad.

Valores de los centroides en la función discriminante

Bajo peso al nacer	Función 1
No	-.267
Si	.588

Tabla 3.22:

Probabilidades A Priori por grupo

Bajo peso al nacer	Pobabilidad A Priori	Casos usados en el análisis	
		No ponderados	Ponderados
No	.688	130	130.000
Si	.312	59	59.000
Total	1.000	189	189.000

Tabla 3.23:

La tabla 3.24 muestra los resultados de la clasificación. Esta tabla es en sí misma un proceso de validación de la función discriminante, pues resume su capacidad predictiva. En ella podemos ver que las madres que no dieron a luz bebés de bajo peso fueron clasificadas correctamente en el 88.5 % de los casos y la madres que sí dieron a luz bebés de bajo peso en el 67.8 % de los casos. En total, la función discriminante consigue clasificar correctamente el 70.9 % de los casos.

Resultados de la clasificación [a]

			Grupo de pertenencia pronosticado		
		Bajo peso al nacer	No	Si	Total
Original	Recuento	No	115	15	130
		Si	40	19	59
	%	No	88.5	11.5	100.0
		Si	67.8	32.2	100.0

a. Clasificados correctamente el 70.9% de los casos agrupados originales

Tabla 3.24:

Conclusiones

Los dos modelos tienen tasas de clasificación correcta muy similares: 73 % de la regresión logística contra 70.9 % del análisis discriminante. En el caso de los bebés con peso normal, la regresión logística pronostica de manera correcta 117 de los 130 casos, mientras que el análisis discriminante lo hace en 115 ocasiones. Con respecto a los bebés que nacen con bajo peso, la regresión logística asigna 21 de los 59 casos en la categoría correcta mientras que el análisis discriminante asigna 19. Si comparamos los resultados que obtenemos con regresión logística y con análisis discriminante, basados en el número de clasificaciones correctas que hace cada uno de ellos, como lo dijimos al inicio de este trabajo, entonces la conclusión general es que la regresión logística arroja los mejores resultados. Estos resultados tienen que ver con el hecho de que el análisis discriminante asume que los datos para las variables representan una muestra proveniente de una distribución normal multivariada, y aunque el no cumplimiento de este supuesto no representa un obstáculo para la realización del análisis, afecta de manera significativa los resultados que se obtienen. Por su parte, la regresión logística no hace ningún supuesto sobre la distribución de las variables, y en ese sentido podemos decir que es un método más robusto comparado con el análisis discriminante.

Siempre que las variables independientes incumplan el supuesto de que están normalmente distribuidas, el uso del análisis discriminante será poco recomendable, pues no existe razón teórica justificada para su suo. La bondad de ajuste será en cierta medida una coincidencia. Sin embargo, cuando los

supuestos se cumplen, la mayor parte de los autores coinciden en que el análisis discriminante arroja mejores resultados que la regresión logística.

En este caso en particular, se podemos concluir que la regresión logística obtiene los mejores resultados, pero en general, la elección depende en gran medida de las necesidades del investigador y su propio enfoque del problema.

En años pasados, unos de los criterios para elegir entre un método y otro era el tiempo de procesamiento de los datos. Como vimos en el capítulo 1, la obtención de los estimadores en la regresión logística se hace mediante métodos iterativos, haciendo que la computación de los resultados fuera mas tardada que en el caso del análisis discriminante. Incluso, era práctica común tomar los estimadores del análisis discriminante como valores de inicio en el proceso iterativo de la regresión logística. Sin embargo, conforme la tecnología va avanzando, estos tiempos de computación se han reducido significativamente, llegando a tomar sólo segundos para obtener los resultados por lo que hoy en día el tiempo no debería ser mas un criterio de decisión para escoger una u otra técnica.

Datos utilizados en el ejercicio del capítulo 3.

ID	LOW	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FTV	BWT
4	1	28	120	3	1	1	0	1	0	709
10	1	29	130	1	0	0	0	1	2	1,021
11	1	34	187	2	1	0	1	0	0	1,135
13	1	25	105	3	0	1	1	0	0	1,330
15	1	25	85	3	0	0	0	1	0	1,474
16	1	27	150	3	0	0	0	0	0	1,588
17	1	23	97	3	0	0	0	1	1	1,588
18	1	24	128	2	0	1	0	0	1	1,701
19	1	24	132	3	0	0	1	0	0	1,729
20	1	21	165	1	1	0	1	0	1	1,790
22	1	32	105	1	1	0	0	0	0	1,818
23	1	19	91	1	1	2	0	1	0	1,885
24	1	25	115	3	0	0	0	0	0	1,893
25	1	16	130	3	0	0	0	0	1	1,899
26	1	25	92	1	1	0	0	0	0	1,928
27	1	20	150	1	1	0	0	0	2	1,928
28	1	21	200	2	0	0	0	1	2	1,928
29	1	24	155	1	1	1	0	0	0	1,936
30	1	21	103	3	0	0	0	0	0	1,970
31	1	20	125	3	0	0	0	1	0	2,055
32	1	25	89	3	0	2	0	0	1	2,055
33	1	19	102	1	0	0	0	0	2	2,082
34	1	19	112	1	1	0	0	1	0	2,084
35	1	26	117	1	1	1	0	0	0	2,084
36	1	24	138	1	0	0	0	0	0	2,100
37	1	17	130	3	1	1	0	1	0	2,125
40	1	20	120	2	1	0	0	0	3	2,126
42	1	22	130	1	1	1	0	1	1	2,187
43	1	27	130	2	0	0	0	1	0	2,187
44	1	20	80	3	1	0	0	1	0	2,211
45	1	17	110	1	1	0	0	0	0	2,225
46	1	25	105	3	0	1	0	0	1	2,240
47	1	20	109	3	0	0	0	0	0	2,240
49	1	18	148	3	0	0	0	0	0	2,282
50	1	18	110	2	1	1	0	0	0	2,296
51	1	20	121	1	1	1	0	1	0	2,296
52	1	21	100	3	0	1	0	0	4	2,301
54	1	26	96	3	0	0	0	0	0	2,325
56	1	31	102	1	1	1	0	0	1	2,353

Tabla 3.25:

57	1	15	110	1	0	0	0	0	0	2,353
59	1	23	187	2	1	0	0	0	1	2,367
60	1	20	122	2	1	0	0	0	0	2,381
61	1	24	105	2	1	0	0	0	0	2,381
62	1	15	115	3	0	0	0	1	0	2,381
63	1	23	120	3	0	0	0	0	0	2,395
65	1	30	142	1	1	1	0	0	0	2,410
67	1	22	130	1	1	0	0	0	1	2,410
68	1	17	120	1	1	0	0	0	3	2,414
69	1	23	110	1	1	1	0	0	0	2,424
71	1	17	120	2	0	0	0	0	2	2,438
75	1	26	154	3	0	1	1	0	1	2,442
76	1	20	105	3	0	0	0	0	3	2,450
77	1	26	190	1	1	0	0	0	0	2,466
78	1	14	101	3	1	1	0	0	0	2,466
79	1	28	95	1	1	0	0	0	2	2,466
81	1	14	100	3	0	0	0	0	2	2,495
82	1	23	94	3	1	0	0	0	0	2,495
83	1	17	142	2	0	0	1	0	0	2,495
84	1	21	130	1	1	0	1	0	3	2,495
85	0	19	182	2	0	0	0	1	0	2,523
86	0	33	155	3	0	0	0	0	3	2,551
87	0	20	105	1	1	0	0	0	1	2,557
88	0	21	108	1	1	0	0	1	2	2,594
89	0	18	107	1	1	0	0	1	0	2,600
91	0	21	124	3	0	0	0	0	0	2,622
92	0	22	118	1	0	0	0	0	1	2,637
93	0	17	103	3	0	0	0	0	1	2,637
94	0	29	123	1	1	0	0	0	1	2,663
95	0	26	113	1	1	0	0	0	0	2,665
96	0	19	95	3	0	0	0	0	0	2,722
97	0	19	150	3	0	0	0	0	1	2,733
98	0	22	95	3	0	0	1	0	0	2,750
99	0	30	107	3	0	1	0	1	2	2,750
100	0	18	100	1	1	0	0	0	0	2,769
101	0	18	100	1	1	0	0	0	0	2,769
102	0	15	98	2	0	0	0	0	0	2,778
103	0	25	118	1	1	0	0	0	3	2,782
104	0	20	120	3	0	0	0	1	0	2,807
105	0	28	120	1	1	0	0	0	1	2,821
106	0	32	121	3	0	0	0	0	2	2,835
107	0	31	100	1	0	0	0	1	3	2,835
108	0	36	202	1	0	0	0	0	1	2,836
109	0	28	120	3	0	0	0	0	0	2,863
111	0	25	120	3	0	0	0	1	2	2,877
112	0	28	167	1	0	0	0	0	0	2,877
113	0	17	122	1	1	0	0	0	0	2,906
114	0	29	150	1	0	0	0	0	2	2,920
115	0	26	168	2	1	0	0	0	0	2,920
116	0	17	113	2	0	0	0	0	1	2,920

Tabla 3.26:

117	0	17	113	2	0	0	0	0	1	2,920
118	0	24	90	1	1	1	0	0	1	2,948
119	0	35	121	2	1	1	0	0	1	2,948
120	0	25	155	1	0	0	0	0	1	2,977
121	0	25	125	2	0	0	0	0	0	2,977
123	0	29	140	1	1	0	0	0	2	2,977
124	0	19	138	1	1	0	0	0	2	2,977
125	0	27	124	1	1	0	0	0	0	2,992
126	0	31	215	1	1	0	0	0	2	3,005
127	0	33	109	1	1	0	0	0	1	3,033
128	0	21	185	2	1	0	0	0	2	3,042
129	0	19	189	1	0	0	0	0	2	3,062
130	0	23	130	2	0	0	0	0	1	3,062
131	0	21	160	1	0	0	0	0	0	3,062
132	0	18	90	1	1	0	0	1	0	3,076
133	0	18	90	1	1	0	0	1	0	3,076
134	0	32	132	1	0	0	0	0	4	3,080
135	0	19	132	3	0	0	0	0	0	3,090
136	0	24	115	1	0	0	0	0	2	3,090
137	0	22	85	3	1	0	0	0	0	3,090
138	0	22	120	1	0	0	1	0	1	3,100
139	0	23	128	3	0	0	0	0	0	3,104
140	0	22	130	1	1	0	0	0	0	3,132
141	0	30	95	1	1	0	0	0	2	3,147
142	0	19	115	3	0	0	0	0	0	3,175
143	0	16	110	3	0	0	0	0	0	3,175
144	0	21	110	3	1	0	0	1	0	3,203
145	0	30	153	3	0	0	0	0	0	3,203
146	0	20	103	3	0	0	0	0	0	3,203
147	0	17	119	3	0	0	0	0	0	3,225
148	0	17	119	3	0	0	0	0	0	3,225
149	0	23	119	3	0	0	0	0	2	3,232
150	0	24	110	3	0	0	0	0	0	3,232
151	0	28	140	1	0	0	0	0	0	3,234
154	0	26	133	3	1	2	0	0	0	3,260
155	0	20	169	3	0	1	0	1	1	3,274
156	0	24	115	3	0	0	0	0	2	3,274
159	0	28	250	3	1	0	0	0	6	3,303
160	0	20	141	1	0	2	0	1	1	3,317
161	0	22	158	2	0	1	0	0	2	3,317
162	0	22	112	1	1	2	0	0	0	3,317
163	0	31	150	3	1	0	0	0	2	3,321
164	0	23	115	3	1	0	0	0	1	3,331
166	0	16	112	2	0	0	0	0	0	3,374
167	0	16	135	1	1	0	0	0	0	3,374
168	0	18	229	2	0	0	0	0	0	3,402
169	0	25	140	1	0	0	0	0	1	3,416
170	0	32	134	1	1	1	0	0	4	3,430
172	0	20	121	2	1	0	0	0	0	3,444
173	0	23	190	1	0	0	0	0	0	3,459

Tabla 3.27:

174	0	22	131	1	0	0	0	0	1	3,460
175	0	32	170	1	0	0	0	0	0	3,473
176	0	30	110	3	0	0	0	0	0	3,475
177	0	20	127	3	0	0	0	0	0	3,487
179	0	23	123	3	0	0	0	0	0	3,544
180	0	17	120	3	1	0	0	0	0	3,572
181	0	19	105	3	0	0	0	0	0	3,572
182	0	23	130	1	0	0	0	0	0	3,586
183	0	36	175	1	0	0	0	0	0	3,600
184	0	22	125	1	0	0	0	0	1	3,614
185	0	24	133	1	0	0	0	0	0	3,614
186	0	21	134	3	0	0	0	0	2	3,629
187	0	19	235	1	1	0	1	0	0	3,629
188	0	25	95	1	1	3	0	1	0	3,637
189	0	16	135	1	1	0	0	0	0	3,643
190	0	29	135	1	0	0	0	0	1	3,651
191	0	29	154	1	0	0	0	0	1	3,651
192	0	19	147	1	1	0	0	0	0	3,651
193	0	19	147	1	1	0	0	0	0	3,651
195	0	30	137	1	0	0	0	0	1	3,699
196	0	24	110	1	0	0	0	0	1	3,728
197	0	19	184	1	1	0	1	0	0	3,756
199	0	24	110	3	0	1	0	0	0	3,770
200	0	23	110	1	0	0	0	0	1	3,770
201	0	20	120	3	0	0	0	0	0	3,770
202	0	25	241	2	0	0	1	0	0	3,790
203	0	30	112	1	0	0	0	0	1	3,799
204	0	22	169	1	0	0	0	0	0	3,827
205	0	18	120	1	1	0	0	0	2	3,856
206	0	16	170	2	0	0	0	0	4	3,860
207	0	32	186	1	0	0	0	0	2	3,860
208	0	18	120	3	0	0	0	0	1	3,884
209	0	29	130	1	1	0	0	0	2	3,884
210	0	33	117	1	0	0	0	1	1	3,912
211	0	20	170	1	1	0	0	0	0	3,940
212	0	28	134	3	0	0	0	0	1	3,941
213	0	14	135	1	0	0	0	0	0	3,941
214	0	28	130	3	0	0	0	0	0	3,969
215	0	25	120	1	0	0	0	0	2	3,983
216	0	16	95	3	0	0	0	0	1	3,997
217	0	20	158	1	0	0	0	0	1	3,997
218	0	26	160	3	0	0	0	0	0	4,054
219	0	21	115	1	0	0	0	0	1	4,054
220	0	22	129	1	0	0	0	0	0	4,111
221	0	25	130	1	0	0	0	0	2	4,153
222	0	31	120	1	0	0	0	0	2	4,167
223	0	35	170	1	0	1	0	0	1	4,174
224	0	19	120	1	1	0	0	0	0	4,238
225	0	24	116	1	0	0	0	0	1	4,593
226	0	45	123	1	0	0	0	0	1	4,990

Tabla 3.28:

Descripción de cada una de las variables utilizadas en el capítulo 3.

Variable	Descripción	Códigos/Valores	Nombre
1	Código de Identificación	ID Number	ID
2	Bajo peso al nacer	1 = BWT<=2500g, 0 = BWT>2500g	LOW
3	Edad de la madre	Años	AGE
4	Peso de la madre en su último periodo menstrual	Libras	LWT
5	Raza	1 = Blanca, 2 = Negra Otra	RACE
6	Status fumador durante el embarazo	0 = No, 1 = Si	SMOKE
7	Historial de parto prematuro	0 = Ninguno, 1 = Uno, 2 = Dos, etc.	PTL
8	Historial de Hipertensión	0 = No, 1 = Si	HT
9	Prescencia de irritabilidad uterina	0 = No, 1 = Si	UI
10	Número de visitas al médico durante el primer trimistre	0 = Ninguno, 1 = Uno 2 = Dos,etc.	FTV
11	Peso al nacer	Gramos	BWT

Tabla 3.29:

Bibliografía

[1] Agresti, A. (1996) *'An introduction to categorical data analysis'* Ed. John Wiley & Sons.

[2] Anderson, R.E., Black, W.C., Hair, J.F. y Tatham, R.L. (1995) *'Multivariate data analysis with readings'*Ed. Prentice Hall.

[3] Hosmer D.W. y Lemeshow S. (1989) *'Applied logistic regression'* Ed. John Wiley & Sons.

[4] Hubert C.J. (1994) *'Aplied Discriminant Analysis'* Ed. John Wiley & Sons.

[5] Jonhson, D.E. (1998) *'Applied multivariate methods for data analysis'*Ed. Brooks Cole Publishing Company.

[6] Jonhson, R.A. y Wichern, D.W. (1992) *'Applied multivariate statistical analysis'* Ed. Prentice Hall.

[7] Press J. y Wilson S. (1978) *'Choosing between logistic regression and discriminant analysis'* Journal of the American Statistical Association

[8] Rencher, A.C. (2002) *'Methods of multivariate analysis'* Ed. John Wiley & Sons.

[9] Simonnoff, J.S. (2003) *'Analyzing categorical data'* Ed. Springer-Verlag.